Numerical Reliability and CPU Time for the Mixed Methods applied to Flow Problems in Porous Media

Hussein Hoteit, Jocelyne Erhel, Robert Mosé, Bernard Philippe, Philippe Ackerer

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Numerical Reliability and CPU Time for the Mixed Methods applied to Flow Problems in Porous Media

Hussein Hoteit — Jocelyne Erhel — Robert Mosé —

Bernard Philippe — Philippe Ackerer

N° 4228

Juillet 2001

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4228--FR+ENG

Numerical Reliability and CPU Time for the Mixed Methods applied to Flow Problems in Porous Media

Hussein Hoteit , Jocelyne Erhel , Robert Mosé *,
Bernard Philippe , Philippe Ackerer *

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Aladin

Rapport de recherche n° 4228 — Juillet 2001 — 43 pages

Abstract: This work is devoted to the numerical reliability and time requirements of the Mixed Finite Element (MFE) and Mixed-Hybrid Finite Element (MHFE) methods. The behavior of these methods is investigated under the influence of two factors: the mesh discretization and the medium heterogeneity. We show that, unlike the MFE, the MHFE "suffers" with the presence of flatted triangular elements. A numerical reliability analyzing software (Aquarels) is used to detect the instability of the matrix-inversion code generated by MAPLE which is used in the MHFE code. We also show that the spectral condition number of the algebraic systems furnished by both methods in heterogeneous media grows up linearly according to the smoothness of the hydraulic conductivity. Furthermore, it is found that the MHFE could accumulate numerical errors if the conductivity varies abruptly in space. Finally, we compare running-times for both algorithms by giving various numerical experiments.

Key-words: elliptic/parabolic problems, flow in porous media, mixed and mixed-hybrid methods, functional stability, symbolic programming.

* Institut de Mécanique des Fluides et des solides, Univ. Louis Pasteur de Strasbourg, CNRS/MUR 7507, 2 rue Boussingault, 67000 Strasbourg.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 Rennes Cedex (France)
Téléphone : +33 2 99 84 71 00 —T élécopie : +33 2 99 84 71 71

Fiabilité numérique et temps CPU requis par les méthodes mixtes appliquées aux problèmes d'écoulement en milieux poreux

Résumé : Ce travail porte sur la fiabilité numérique et le temps d'exécution des méthodes des Eléments Finis Mixtes (EFM) et des Eléments Finis Mixtes Hybrides (EFMH). Le comportement de ces méthodes est étudié sous l'influence de deux facteurs: la discrétisation spatiale et l'hétérogénéité du milieu. Nous prouvons que, contrairement à la méthode des EFM, la méthode des EFMH "souffre" en présence d'éléments triangulaires très aplatis. Un atelier pour la fiabilité numérique (Aquarels) est employé pour détecter l'instabilité du code d'inversion de matrice généré par MAPLE qui est utilisé dans le code EFMH. Nous prouvons également que le conditionnement des systèmes algébriques fournis par les deux méthodes dans des milieux hétérogènes croit linéairement selon le contraste des conductivités hydrauliques. En outre, nous observons que la méthode des EFMH peut accumuler des erreurs numériques inacceptables si la conductivité varie brutalement en espace. Enfin, nous comparons le temps d'exécution des deux algorithmes en effectuant diverses expériences numériques.

Mots-clés : problèmes elliptiques/paraboliques, écoulement en milieu poreux, éléments finis mixtes et mixtes hybrides, stabilité fonctionnelle, programmation symbolique.

1 Introduction

Various transient problems in the science and engineering fields, such as the heat transfer, electromagnetic current, flow of fluids and transport of solute in porous media [1] etc., are governed by coupled systems of time-dependent partial differential equations. Due to the powerlessness of classical methods like the finite element or finite difference methods in manipulating these systems where usually the primary variable and its derivative are needed to be approximated, the mixed and mixed-hybrid finite element methods are developed to handle such problems. The main favorable property of these methods is that both together the primary unknown and its gradient are approximated simultaneously with the same order of convergence. Besides, they fulfill the physics of the problem, i.e. conserve mass locally and preserve the continuity of fluxes (see, e.g., [4, 5]).

The foremost motivation of this work is to give a scrupulous examination of the numerical reliability and time-consuming of the MFE and MHFE methods applied to elliptic/parabolic problems. The Darcy's law and the mass conservation equation describing a single phase fluid flow in porous media will be studied. In the case of transient flow, the parabolic governing equations for the unknown pressure head scalar function p and Darcy's velocity vector function u are given by

$$\begin{aligned} s(x)\,\frac{\partial p(x,t)}{\partial t} + \nabla.u(x,t) &= f(x,t) && \text{in} && \Omega\times(0,T],\\ u(x,t) &= -\mathcal{K}(x)\,\nabla p(x,t) && \text{in} && \Omega\times(0,T],\\ p(x,0) &= p^0(x) && \text{in} && \Omega,\\ p(x,t) &= p^D(x,t) && \text{on} && \Gamma^D\times(0,T],\\ u(x,t).\nu &= q^N(x,t) && \text{on} && \Gamma^N\times(0,T]. \end{aligned} \tag{1}$$

In the case of steady flow, the stationary problem of (1) is reduced to the following second order elliptic equations

$$\begin{aligned} \nabla.u(x) &= f(x) && \text{in} && \Omega,\\ u(x) &= -\mathcal{K}(x)\,\nabla p(x) && \text{in} && \Omega,\\ p(x) &= p^D(x) && \text{on} && \Gamma^D,\\ u(x).\nu &= q^N(x) && \text{on} && \Gamma^N, \end{aligned} \tag{2}$$

where Ω is a bounded domain in R^d $(d = 1,\ldots,3)$ with boundary $\partial\Omega = \Gamma^D \cup \Gamma^N$; $\mathcal{K} = \mathcal{K}(x)$ is the so-called hydraulic conductivity (permeability), it is assumed to be a diagonal tensor with components in $L^\infty(\Omega)$; ν indicates the outward unit normal

vector along $\partial\Omega$; $f \in L^2(\Omega)$ represents the sink/source function; s is the storage coefficient; p^D and q^N are respectively the Dirichlet and Neumann boundary conditions.

It is well known that the MFE formulation, in approximating the stationary problem (2), leads to a saddle point problem (see, e.g., [3, 4, 6, 9]). Therefore, an indefinite algebraic linear system is obtained and consequently cannot be solved by direct usage of robust algorithms like Cholesky or Conjugate Gradient methods. The hybridization idea is exerted to overcome this problem, hereby new degrees of freedom are appended. The benefit of this approach is that it leads to solve an equivalent linear system which is symmetric and positive definite. It was shown in [6, 10] that by using the lumped-mass technique the mixed formulation with rectangular elements boils down to the classical cell-centered finite differences. Nevertheless, in this work, we study the mixed formulations by using analytic integrations and without any restrictions on the discretized elements. It is found that, unlike the circumstances in the case of elliptic problems, the MFE method leads to solve a symmetric, positive definite linear system in approximating pure transient parabolic problems. Furthermore, we show that the MFE method is numerically more accurate than the MHFE in approximating the fluxes, chiefly with the presence of flat mesh elements or large variations in the medium heterogeneity. By discretizing the domain into triangular elements, the mixed-hybrid formulation necessitates inverting a 3×3–dimensional matrix for each element. It is found that flat triangles could blow up the conditioning of the corresponding matrices, so one should be cautious in the way whereby these matrices are inverted. Incipiently, by using a matrix-inversion subroutine automatically generated by Maple led sometimes to non-consistent results that are mostly obtained on relatively flat elements. The matrix-inversion function of Maple is based on Cramer's rule which is well known to be numerically unstable [8]. Accordingly, the instability of the Maple's subroutine is shown by using a numerical stability detecting software, Aquarels [18, 19]. Comparisons with another matrix-inversion code based on LDL–factorization method show that this code is stable, besides it is more efficient than the former. On the other hand, the surpassing property of the MFE method is that the inversion of such matrices is avoidable.

Generally, rough physical parameters in heterogeneous media could cause shortcomings in the approximated solutions. The weak spot of classical methods is that the velocity unknown u is approximated by numerical differentiation of the primary

unknown and then multiplication by an often rough tensor of conductivity. In the works presented in [11, 13], numerous numerical experiments attest that the MHFE method is numerically more reliable than the conforming finite element method. However, in this work we inspect the behavior of the algebraic systems and the resulting solutions fulfilled by both mixed methods. We prove that the conditioning of the resulting algebraic linear systems grows up linearly according to the ratio between the highest and lowest values of the hydraulic conductivity of adjacent elements in heterogeneous media. Furthermore, we have detected that the MHFE algorithm could accumulate numerical errors if large jumps in the tensor of conductivity take place. In accordance with the work presented in [6], it is found that the condition number of the linear system induced by the MFE method is critically affected by the values of the storage coefficient s.

This paper is organized as follows. In the next section, after reviewing the approximation spaces and the variational formulations of the equations (1), (2), we present the elementary equations and the final algebraic systems derived from both mixed methods. The aim in section 3 is to study the effect of the mesh geometry on the approximated solutions. Thereat, the numerical reliability of two matrix-inversion subroutines is analyzed by using Aquarels. The properties of the algebraic systems induced by both methods in simulating flow in heterogeneous media are investigated in section 4. Finally, before ending with a conclusion, we give in section 5 some numerical experiments comparing running-times of the MHFE and MFE algorithms.

2 Mixed and Mixed-Hybrid Finite Element Discretizations

The essential idea of the mixed methods is to approximate individually the Darcy's law and flow equation and we get additionally the Darcy velocity u as an unknown function. Thus, the variation formulations of the given PDEs systems are chosen in a way to have the pressure and its gradient in the basic formulation.
Introducing the Hilbert spaces

$$\begin{aligned} H(div;\Omega) &= \{\chi \in (L^2(\Omega))^2 \mid \nabla.\chi \in L^2(\Omega)\}, \\ H_{g,N}(div;\Omega) &= \{\chi \in H(div;\Omega) \mid \nu.\chi = g \text{ on } \Gamma^N\} \end{aligned}$$

the mixed formulation of (1) can be stated as:

Find $(u,p) \in H_{q^N,N}(div;\Omega) \times L_2(\Omega)$, such that

$$\left\{ \begin{array}{rrl} \displaystyle\int_\Omega (\mathcal{K}^{-1}\, u).\chi \, dx + \int_{\partial\Omega} p^D \, \nu.\chi \, d\ell & = \displaystyle\int_\Omega p \nabla.\chi \, dx & \forall\, \chi \in H_{0,N}(div;\Omega), \\ \displaystyle\int_\Omega s\frac{\partial p}{\partial t}\, \varphi \, dx + \int_\Omega \nabla.u \, \varphi \, dx & = \displaystyle\int_\Omega f \, \varphi \, dx & \forall\, \varphi \in L_2(\Omega). \end{array} \right. \tag{3}$$

In order to state a finite element formulation of problem (3) it is necessary to define finite dimensional subspaces of $H(div;\Omega)$ and $L_2(\Omega)$. These spaces are, in the simplest case, the Raviart-Thomas spaces of lowest order RT^0 and multiplier spaces $\mathcal{M}^0$.
We shall restrict our discussion to the case of two-dimensional triangular discretization. The other spatial discretizations follow in a similar manner. Throughout this paper, we denote by $\mathcal{T}_h$ the set of triangular partitioned elements of Ω where h refers to the maximal mesh spacing ($h = \max\limits_{K\in\mathcal{T}_h} diam(K)$). Let $\mathcal{E}_h$ be the collection of edges of the grid not belonging to Γ^D. By $N_\mathcal{E}$ and $N_\mathcal{T}$ we denote the cardinals of $\mathcal{E}_h$ and $\mathcal{T}_h$, respectively.
Define the Raviart-Thomas spaces

$$\begin{array}{lcl} RT^0(K) & = & \{s \in (\mathcal{P}_1(K))^2 \mid s = (a + b\,x_1, c + b\,x_2),\ a,b,c \in \mathbb{R}\}, \\ RT^0(\mathcal{T}_h) & = & \{\phi \in L^2(\Omega) \mid \phi/_K \in RT^0(K)\ \ \forall K \in \mathcal{T}_h\}, \\ RT^0_{g,N}(\mathcal{T}_h) & = & RT^0(\mathcal{T}_h) \cap H_{g,N}(div;\Omega), \end{array}$$

where $\mathcal{P}_d(K)$ is the space of polynomials of total degree d defined on K.
Further, the multiplier space $\mathcal{M}^0(\mathcal{T}_h)$ is defined as

$$\mathcal{M}^0(\mathcal{T}_h) = \{\varphi \in L^2(\Omega) \mid \varphi/_K \in \mathcal{P}_0(K), K \in \mathcal{T}_h\}.$$

The lowest order Raviart-Thomas mixed discretization of problem (3) reads as follows:

Find $(u_h, p_h) \in RT^0_{q^N,N}(\mathcal{T}_h) \times \mathcal{M}^0(\mathcal{T}_h)$, such that

$$\left\{ \begin{array}{rrl} \displaystyle\int_\Omega (\mathcal{K}^{-1}\, u_h).\chi_h \, dx + \int_{\partial\Omega} p^D \, \nu.\chi_h \, d\ell & = \displaystyle\int_\Omega p_h \nabla.\chi_h \, dx & \forall\, \chi_h \in RT^0_{0,N}(\mathcal{T}_h), \\ \displaystyle\int_\Omega s\frac{\partial p_h}{\partial t}\, \varphi_h \, dx + \int_\Omega \nabla.u_h \, \varphi_h \, dx & = \displaystyle\int_\Omega f \, \varphi_h \, dx & \forall\, \varphi_h \in \mathcal{M}^0(\mathcal{T}_h). \end{array} \right. \tag{4}$$

On the other hand, in the MHFE formulation, u_h is sought in the enlarged Raviart-Thomas space $RT^0(\mathcal{T}_h)$. The continuity of the normal flux across the interelement boundaries is enforced by Lagrange multipliers on the space of constant functions $\mathcal{N}^0(\mathcal{E}_h)$ over the edges. Define the multiplier spaces

$$\begin{array}{rcl} \mathcal{N}^0(\mathcal{E}_h) & = & \{\lambda \in L^2(\mathcal{E}_h) \mid \lambda/_E \in \mathcal{P}_0(E) \ \ \forall E \in \mathcal{E}_h\}, \\ \mathcal{N}^0_{g,D}(\mathcal{E}_h) & = & \{\lambda \in \mathcal{N}^0(\mathcal{E}_h) \mid \lambda = g \text{ on } \Gamma^D\}. \end{array}$$

Then the mixed hybrid discretization reads as:

Find $(u_h, p_h, tp_h) \in RT^0(\mathcal{T}_h) \times \mathcal{M}^0(\mathcal{T}_h) \times \mathcal{N}^0_{p^D,D}(\mathcal{E}_h)$ such that

$$\left\{ \begin{array}{ll} \displaystyle\int_\Omega (\mathcal{K}^{-1} u_h).\chi_h \, dx + \sum_{K\in\mathcal{T}_h} \int_{\partial K} tp_h \, \nu_K.\chi_h \, d\ell = \sum_{K\in\mathcal{T}_h} \int_K p_h \nabla.\chi_h \, dx & \forall \, \chi_h \in RT^0(\mathcal{T}_h), \\ \displaystyle\int_\Omega s\frac{\partial p_h}{\partial t} \, \varphi_h \, dx + \int_\Omega \nabla.u_h \, \varphi_h \, dx = \int_\Omega f \, \varphi_h \, dx & \forall \, \varphi_h \in \mathcal{M}^0(\mathcal{T}_h), \\ \displaystyle\sum_{K\in\mathcal{T}_h} \int_{\partial K} u_h.\nu_K \, \lambda_h \, d\ell = \int_{\partial\Omega} q^N \, \lambda_h \, d\ell & \forall \, \lambda_h \in \mathcal{N}^0_{0,D}(\mathcal{E}_h). \end{array} \right. \tag{5}$$

2.1 Local basis functions

The Raviart-Thomas basis functions of the 3-dimensional space $RT^0(K)$ are defined as

$$w_{K,E_i} = \frac{1}{2|K|} \begin{pmatrix} x_1 - x_{1i} \\ x_2 - x_{2i} \end{pmatrix} \qquad i = 1, \ldots, 3, \tag{6}$$

where $|K|$ is the measure of the triangular element K and the (x_{1i}, x_{2i})'s are its vertices (see Fig. 1).
Therefore, for every $\chi_K \in RT^0(K)$, $K \in \mathcal{T}_h$, it can be written as $\chi_K = \displaystyle\sum_{E\subset\partial K} q_{K,E} \, w_{K,E}$.
Furthermore, the following properties are satisfied.

i. $\nabla.\chi_K$ is constant over K.

ii. $\displaystyle\int_E \nu_{K,E}.\chi_K \, d\ell = q_{K,E}$ is constant on each $E \subset \partial K$.

Hence, u_K is uniquely determined by the normal fluxes $q_{K,E}$ across the edges of K, where $\nu_{K,E}$ denotes the outer normal vector on E with respect to K.

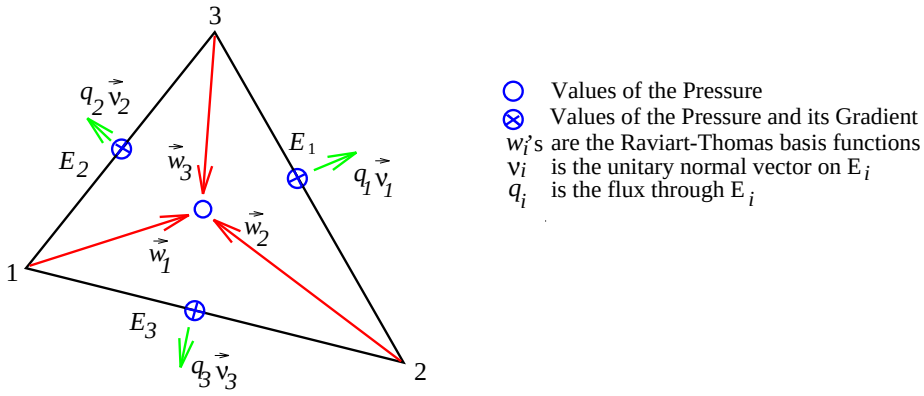

Figure 1: Nodal points and basis functions on triangular elements.

2.2 Mixed-Hybrid Finite Element formulation

The finite dimensional space $RT^0(\mathcal{T}_h)$ is spanned by linearly independent vectorial basis functions $w_{K,E}$, $E \subset \partial K$, $K \in \mathcal{T}_h$, such that $w_{K,E}$ has its support in K ($supp(w_{K,E}) \subseteq K$) and

$$\int_{E'} w_{K,E}.\nu_K \, d\ell = \delta_{EE'}, \quad E, E' \subset \partial K.$$

These functions can be chosen the local bases functions given in (6). Thus, a function $u_h \in RT^0(\mathcal{T}_h)$ has three degrees of freedom per element which are the fluxes across the element's edges

$$u_h(x) = \sum_{K \in \mathcal{T}_h} \sum_{E \subset \partial K} q_{K,E} \, w_{K,E}(x), \; x \in \Omega.$$

The two spaces $\mathcal{M}^0(\mathcal{T}_h)$ and $\mathcal{N}^0(\mathcal{E}_h)$ are spanned respectively by the linearly independent scalar basis functions φ_K, $K \in (\mathcal{T}_h)$, and λ_E, $E \in (\mathcal{E}_h)$, such that

$$\begin{aligned} \varphi_K(x) &= \delta_{K,K'}, \; x \in K', \quad K, K' \in \mathcal{T}_h, \\ \lambda_E(x) &= \delta_{E,E'}, \; x \in E', \quad E, E' \in \mathcal{E}_h. \end{aligned}$$

Thus, a function $p_h \in \mathcal{M}^0(\mathcal{T}_h)$ (resp. $tp_h \in \mathcal{N}^0(\mathcal{E}_h)$) has one degree of freedom of constant value per element $K \in \mathcal{T}_h$ (resp. $E \in \mathcal{E}_h$), such that

$$\begin{aligned} p_h(x) &= \sum_{K' \in \mathcal{T}_h} p_{K'}\, \varphi_{K'}(x) = p_K, \quad x \in K, \\ tp_h(x) &= \sum_{E' \in \mathcal{E}_h} tp_{E'}\, \lambda_{E'}(x) = tp_E, \quad x \in E. \end{aligned}$$

Now, we individually investigate the underlying equations in (5), which can be integrated over the element level.

2.2.1 Discretization of Darcy's law

By taking as test functions χ_K successively the basis functions $w_{K,E}$, the discretized equation of Darcy's law (the first equation in 5) becomes

$$\int_K (\mathcal{K}_K^{-1}\, u_K).\chi_K dx + \sum_{E \subset \partial K} \int_E tp_{K,E}\, \chi_K.\nu_{K,E} d\ell = \int_K p_K \nabla.\chi_K dx, \tag{7}$$

where $\mathcal{K}_K$ is a piecewise approximation of the conductivity tensor over K, and

$$tp_E = tp_{K,E} = \begin{cases} tp_{K',E} & if\ E = K \cap K' \\ p_E^D & if\ E \in \Gamma^D \end{cases}, \quad E \in \mathcal{E}_h \cup \Gamma^D, \quad K,\, K' \in \mathcal{T}_h.$$

By integrating (7) and by making use of the Raviart-Thomas space basis properties, the following equations come into view

$$\sum_{E' \subset \partial K} (B_K)_{E,E'}\, q_{K,E'} = p_K - tp_{K,E}, \qquad E \subset \partial K,\, K \in \mathcal{T}_h. \tag{8}$$

They can be written in the matrix form

$$B_K\, Q_K = p_K\, e - T_{P_K}, \qquad K \in \mathcal{T}_h, \tag{9}$$

where

Q_K and T_{P_K} are 3– dimensional vectors containing respectively the fluxes $q_{K,E}$ and the traces of the pressure $tp_{K,E}$ on each $E \subset \partial K$;

e refers to the elementary divergence vector. It is of dimension 3 and unitary entries;

B_K is a 3×3 symmetric positive definite matrix whose elements are

$$(B_K)_{E,E'} = \int_K w_{K,E}^T \, \mathcal{K}_K^{-1} \, w_{K,E'} \, dx. \tag{10}$$

It should be noted that these integrations are all evaluated exactly.

The last equation in (5) is equivalent to

$$\begin{aligned} &\int_E u_K.\nu_{K,E} \, d\ell + \int_E u_{K'}.\nu_{K',E} \, d\ell = 0 \qquad if \quad E = K \cap K', \\ &\int_E u_K.\nu_{K,E} \, d\ell = q_E^N \qquad\qquad\qquad\qquad if \quad E \in \Gamma^N, \end{aligned}$$

where $q_E^N = \int_E q^N \, d\ell$.
Hence, the normal components of u_h are continuous across the interelement boundaries, i.e.

$$q_{K,E} = \begin{cases} -q_{K',E} & if \; E = K \cap K', \\ q_E^N & if \; E \in \Gamma^N. \end{cases} \tag{11}$$

By inverting the matrix B_K and using (11), it is possible to eliminate the unknown flux (in the runs we shall investigate in details the way whereby these matrices are inverted). As a result, the reduced algebraic system, acquired by discretizing Darcy's law with unknowns the pressure head given in P and its traces in T_P, becomes

$$R^T \, P - M \, T_P + V = 0, \tag{12}$$

where

R^T is the transpose matrix of R which is a sparse matrix of dimension $N_\mathcal{E} \times N_\mathcal{T}$ with nonzero elements given by

$$(R)_{K,E} = \alpha_{K,E} = \sum_{E' \subset \partial K} (B_K^{-1})_{E,E'} \, , \qquad E \subset \partial K;$$

M is a $N_\mathcal{E} \times N_\mathcal{E}$ sparse matrix with nonzero entries defined as

$$(M)_{E,E'} = \sum_{\partial K \supset E,E'} (B_K^{-1})_{E,E'};$$

V is a $N_\mathcal{E}$–dimensional vector corresponding to the Dirichlet and Neumann boundary conditions.

2.2.2 Discretization of the mass conservation equation

By integrating the mass conservation equation (the second equation in (5)) where the test functions ϕ_h are successively replaced by the basis functions of $\mathcal{M}^0$, we get

$$s_{_K}|K|\,\frac{\partial p_{_K}}{\partial t} + \sum_{E\subset K} q_{K,E} = f_{_K} \qquad K \in \mathcal{T}_h, \tag{13}$$

where $s_{_K}$ and $f_{_K}$ are respectively the approximations of the storage coefficient and the sink/source term over K.
Therefrom, by using (9) to replace the sum of fluxes in (13), we obtain an ordinary differential system which is given in its matrix form

$$S\,\frac{dP}{dt} + D\,P - R\,T_P = F, \tag{14}$$

where

S is a $N_{\mathcal{T}} \times N_{\mathcal{T}}$ diagonal matrix with entries $(S)_{_{K,K}} = s_{_K}\,|K|\,;$

D is also a $N_{\mathcal{T}} \times N_{\mathcal{T}}$ diagonal matrix whose coefficients are

$$(D)_{_{K,K}} = \alpha_{_K} = \sum_{E\subset\partial K} \alpha_{K,E}\,;$$

F is a vector of dimension $N_{\mathcal{T}}$, it corresponds to the source/sink function as well as to the imposed pressure given by the Dirichlet boundary conditions.

2.2.3 The derived algebraic systems

The spatial discretization of the governing equations obtained by applying the mixed-hybrid formulation led to two systems. The first one, given in (12), is an algebraic system of unknowns P and T_P and the second is an ordinary system of first order differential equations in time (14). They can be written in the matrix form

$$\begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{dP}{dt} \\ \frac{dT_P}{dt} \end{pmatrix} + \begin{pmatrix} D & -R \\ -R^T & M \end{pmatrix} \begin{pmatrix} P \\ T_P \end{pmatrix} = \begin{pmatrix} F \\ V \end{pmatrix}. \tag{15}$$

Due to the fact that using exact time integrations in solving (15) is computationally very consuming [14], a temporal discretization of the differential operator is required. Using for the sake of simplicity the first-order backward Euler (implicit) scheme,

denoting by Δt the sampling time-step and using the superscript n to refer to the nth time level, we obtain the following system for each $n > 0$

$$(\mathcal{L} + \Delta t\, \mathcal{J}) \begin{pmatrix} P^n \\ T_P^n \end{pmatrix} = \mathcal{L} \begin{pmatrix} P^{n-1} \\ T_P^{n-1} \end{pmatrix} + \Delta t \begin{pmatrix} F^n \\ V^n \end{pmatrix}, \tag{16}$$

where $\mathcal{L} = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$, $\mathcal{J} = \begin{pmatrix} D & -R \\ -R^T & M \end{pmatrix}$.

2.2.4 Properties of the algebraic systems

Here, we present some properties of the algebraic systems induced by the MHFE formulation.

Lemma 2.1 *For any triangular element K, the elementary matrix B_K has $e = (1\,1\,1)^T$ as an eigenvector with $3/\alpha_K$ the corresponding eigenvalue. Moreover,*

$$\alpha_{K,E} = \alpha_{K,E'} = 1/3\,\alpha_K, \quad \forall E,\, E' \subset \partial K.$$

Proof:

Let K be any element in $\mathcal{T}_h$ (see Fig.1) with vertices (x_{1i}, x_{2i}) and edges E_i, $i = 1, \ldots, 3$. We denote by $(\bar{x}_1, \bar{x}_2)$ the barycenter of K.

Let $\beta_{K,E} = \sum_{E' \subset \partial K} (B_K)_{E,E'}$, $\forall\, E,\, E' \in \partial K$. By using the shape functions given in (6), we have

$$\begin{aligned}
\beta_{K,E_i} &= \sum_{\ell=1}^{3} \int_K w_{K,E_i}^T\, \mathcal{K}_K^{-1}\, w_{K,E_\ell}\, dx \\
&= \int_K w_{K,E_i}^T\, \mathcal{K}_K^{-1} \sum_{\ell=1}^{3} w_{K,E_\ell}\, dx \\
&= \frac{3}{2|K|} \int_K (x_1 - x_{1i} \;\; x_2 - x_{2i}) \mathcal{K}_K^{-1} (x_1 - \bar{x}_1 \;\; x_2 - \bar{x}_2)^T\, dx \quad \text{for } i = 1, \ldots, 3.
\end{aligned}$$

Then $\forall\, i, j \in \{1, 2, 3\}, i \neq j$, we finally obtain

$$\begin{aligned}
\beta_{K,E_i} - \beta_{K,E_j} &= \frac{3}{2|K|} \int_K (x_{1j} - x_{1i} \;\; x_{2j} - x_{2i})\, \mathcal{K}_K^{-1}\, (x_1 - \bar{x}_1 \;\; x_2 - \bar{x}_2)^T\, dx \\
&= \frac{3}{2|K|}\, (x_{1j} - x_{1i} \;\; x_{2j} - x_{2i})\, \mathcal{K}_K^{-1} \left(\int_K (x_1 - \bar{x}_1)\, dx \;\; \int_K (x_2 - \bar{x}_2)\, dx \right)^T \\
&= 0.
\end{aligned}$$

One can easily verify that $\int_K (x_1 - \bar{x}_1)\,dx = \int_K (x_2 - \bar{x}_2)\,dx = 0$.
Thus, e is an eigenvector of B_K and consequently it is also an eigenvector of B_K^{-1}.

□

Lemma 2.2 *The matrix M is symmetric, positive definite.*

Proof:
For any nonzero $y \in \mathbb{R}^{N_\mathcal{E}}$, we have

$$\begin{aligned} y^T M y &= \sum_{E,E' \in \mathcal{E}_h} y_E M_{E,E'} y_{E'} \\ &= \sum_{K \in \mathcal{T}_h} \sum_{\partial K \supset E,E'} y_{K,E} (B_K^{-1})_{E,E'} y_{K,E'} \\ &= \sum_{K \in \mathcal{T}_h} y_K^T (B_K^{-1})\, y_K, \end{aligned}$$

where $y_K = (y_{K,E})_{E \subset \partial K} \in \mathbb{R}^3$.
From lemma (2.1), B_K^{-1} is positive definite and so is M.

□

Proposition 2.1 *With the presence of Dirichlet boundary conditions, the matrix*

$$\mathcal{J} = \begin{pmatrix} D & -R \\ -R^T & M \end{pmatrix}$$

is positive definite, otherwise it is semidefinite.

Proof:
For any nonzero vector $(x\ y) \in \mathbb{R}^{N_\mathcal{T}} \times \mathbb{R}^{N_\mathcal{E}}$, we have

$$(x^T\ y^T)\mathcal{J} \begin{pmatrix} x \\ y \end{pmatrix} = x^T D\, x - 2\, x^T R\, y + y^T M y.$$

D is a diagonal matrix, then

$$x^T D\, x = \sum_{K \in \mathcal{T}_h} x_K^2\, \alpha_K, \quad \text{where } x = (x_K)_{N_\mathcal{T}}.$$

Let $e = (1\,1\,1)^T$ and $y = (y_E)_{N_\mathcal{E}} = (y_K^T)_{N_\mathcal{T}}$ with $y_K = (y_{K,E})_{E\subset\partial K} \in \mathbb{R}^3$, then we get

$$\begin{aligned} x^T R\, y &= \sum_{K\in\mathcal{T}_h} \sum_{E\subset\mathcal{E}_h} x_K R_{K,E} y_E \\ &= \sum_{K\in\mathcal{T}_h} \sum_{E\subset\partial K} x_K \alpha_{K,E}\; y_{K,E} \\ &= \sum_{K\in\mathcal{T}_h} x_K \frac{\alpha_K}{3}\; e^T y_K. \end{aligned}$$

We set $z_K = y_K - \zeta_K\, e$ with $\zeta_K = \sum_{E\subset\partial K} y_{K,E}$, then

$$\begin{aligned} y^T M\, y &= \sum_{K\in\mathcal{T}_h} y_K^T (B_K^{-1})\, y_K \\ &= \sum_{K\in\mathcal{T}_h} \left[\alpha_K \zeta_K^2 + z_K^T (B_K^{-1})\, z_K\right]. \end{aligned}$$

By simple calculations, we get

$$(x^T\; y^T)\mathcal{J}\begin{pmatrix} x \\ y \end{pmatrix} = \sum_{K\in\mathcal{T}_h} \left[\alpha_K (x_K - \zeta_K)^2 + z_K^T (B_K^{-1})\, z_K\right]. \tag{17}$$

Thus, (17) is strictly positive whenever $z_K \neq 0$ or $x_K \neq \zeta_K$ for some $K \in \mathcal{T}_h$. Therefore, $\mathcal{J}$ is positive semidefinite.

Now, suppose that $\Gamma^D \neq \emptyset$ so there exists a boundary element $K^D \in \mathcal{T}_h$ and $E^D \subset \partial K$, such that $E^D \in \Gamma^D$. We will prove that $\mathcal{J}$ becomes definite positive.
By taking $z_K = 0$ and $x_K = \zeta_K$, $\forall K \in \mathcal{T}_h$, $K \neq K^D$, (17) is reduced to

$$(x^T\; y^T)\mathcal{J}\begin{pmatrix} x \\ y \end{pmatrix} = x_{K^D}^2\; \alpha_{K^D} - 2\, x_{K^D} \frac{\alpha_{K^D}}{3}\; e^T y_{K^D} + y_{K^D}^T (B_{K^D}^{-1})\, y_{K^D}.$$

We have

$$y_{K^D,E} = \begin{cases} x_{K^D} & if\ E \neq E^D, \\ 0 & otherwise, \end{cases}$$

since the row and the column corresponding to the imposed edge (E^D) are eliminated from $\mathcal{J}$. Thus, one can easily verify that

$$y_{K^D}^T (B_{K^D}^{-1})\, y_{K^D} = x_{K^D}^2 \left(\frac{\alpha_{K^D}}{3} + (B_{K^D}^{-1})_{E^D,E^D} \right).$$

Finally, we get

$$\begin{aligned} (x^T\ y^T)\mathcal{J} \begin{pmatrix} x \\ y \end{pmatrix} &= x_{K^D}^2\ \alpha_{K^D} - 4\, x_{K^D}^2\, \frac{\alpha_{K^D}}{3} + x_{K^D}^2 \left(\frac{\alpha_{K^D}}{3} + (B_{K^D}^{-1})_{E^D,E^D} \right) \\ &= x_{K^D}^2\ (B_{K^D}^{-1})_{E^D,E^D} > 0. \end{aligned}$$

Hence, $\mathcal{J}$ is positive definite.

□

Corollary 2.1 *The matrix*

$$(\mathcal{L} + \Delta t\, \mathcal{J}) = \begin{pmatrix} S + \Delta t\, D & -\Delta t\, R \\ -\Delta t\, R^T & \Delta t\, M \end{pmatrix}$$

is positive definite.

Proof:

Let $z = (x\ y)^T$ be a nonzero vector in $\mathbb{R}^{N_\mathcal{T}} \times \mathbb{R}^{N_\mathcal{E}}$. The two matrices S and $\mathcal{J}$ are positive definite and semidefinite matrices, respectively. Thus, we get

$$z^T(\mathcal{L} + \Delta t \mathcal{J})z = x^T S\, x + \Delta t\, z^T \mathcal{J} z > 0,$$

since by taking $x = 0$, the quantity $z^T \mathcal{J} z$ is strictly positive (see the proof of (2.1).

□

Since the matrix $G = (S + \Delta t\, D)$ is diagonal, it can be easily inverted. Hence, by eliminating P^n from (16), the following Schur complement system is obtained

$$\begin{cases} (M - \Delta t\, R^T G^{-1} R)\, T_P^n = R^T\, G^{-1}\, (S\, P^{n-1} + \Delta t F^n) + V^n. \\ G\, P^n = S\, P^{n-1} + \Delta t\, R\, T_P^n + \Delta t\, F^n. \end{cases} \tag{18}$$

Proposition 2.2 *The Schur complement matrix* $(M - \Delta t\, R^T G^{-1} R)$ *is positive definite.*

Proof:

let $y \in \mathbb{R}^{N_\mathcal{E}}$ be a nonzero vector, then

$$\begin{aligned} y^T(M - \Delta t\, R^T G^{-1} R)y &= y^T M y - \Delta t\, y^T R^T G^{-1} R\, y \\ &= \frac{1}{\Delta t}(x^T y^T)(\mathcal{L} + \Delta t\, \mathcal{J}) \begin{pmatrix} x \\ y \end{pmatrix} > 0, \end{aligned}$$

where x is chosen to be $\Delta t\, G^{-1} R\, y$.
Therefore, our proposition holds by applying corollary (2.1).

□

As a result, the MHFE formulation leads to compute, at every time step, first T_P by solving a linear system with symmetric, positive definite coefficient matrix, then P by solving a diagonal linear system. As a matter of fact, experimental inspections showed the adaptability and the robustness of the preconditioned conjugate gradient method in solving such systems [11].

The principal steps of the MHFE algorithm can be illustrated as follows.

Algorithm 1 *Principal steps of the MHFE algorithm.*

1– Initialize geometry and physical parameters of the problem.
2– Create the Schur complement matrix.
3– *Iterations on the time–steps.*
  4– Find T_P *by solving the first system in (18).*
  5– Find P *by solving the second system in (18).*
  6– Loop on the number of cells.
    7– Evaluate and invert B_K*.*
    8– Calculate the flux Q_K *by solving (9).*
  9– Write the outputs P *and* Q*.*

It should be noted that at each time step we have to invert B_K (as appears in Algo. 1) which could be very time-consuming. On the other hand, if one stores the matrices B_K^{-1} for each element K, this will relatively exhaust the memory capacities.

2.2.5 Discretization of the time independent problem

In the approximation of the time independent problem (2), the derived algebraic system (15) is reduced to

$$\mathcal{J}\begin{pmatrix} P \\ T_P \end{pmatrix} = \begin{pmatrix} F \\ V \end{pmatrix}. \tag{19}$$

By inverting D, the Schur complement system becomes

$$\begin{cases} (M - R^T D^{-1} R)\, T_P = R^T D^{-1} F + V, \\ DP = R\, T_P + F. \end{cases} \tag{20}$$

Once again, by using proposition (2.1), one can easily show the positive definitiveness of the Schur complement matrix $(M - R^T D^{-1} R)$. Furthermore, it is found that this system is efficiently solved by the preconditioned conjugate gradient method [11, 15].

2.3 Mixed Finite Element formulation

The last equation in (5) ensures the continuity of the normal components of u_h across the interelement boundaries, i.e. $u_h \in RT^0_{p^D,N}(\mathcal{T}_h)$. Therefore, if $(u_h, p_h, tp_h) \in RT^0(\mathcal{T}_h) \times \mathcal{M}^0(\mathcal{T}_h) \times \mathcal{N}^0_{p^D,D}(\mathcal{E}_h)$ is the solution of (5) then (u_h, p_h) is also the solution of (4) (see, e.g., [3]). Thus, the two methods are in fact two different formulations of the same numerical approximation. Hence, the MFE solution (u_h, p_h) can be simply deduced from that of the MHFE by eliminating the pressure traces tp_h and taking as main unknowns the pressure and the fluxes across the mesh edges.
In order to fulfill the pressure and flux continuity constraints, we introduce a scalar sign indicator $\epsilon_{K,E}$ similar to that used in [3, 6],

$$\epsilon_{K,E} = \begin{cases} \nu_K . \nu_{K,E} & \text{if } E \subset \partial K, \\ 0 & \text{if } E \not\subset \partial K, \end{cases} \tag{21}$$

where ν_E is an arbitrary chosen unitary normal vector on E, and $\nu_{K,E}$ is the outer unitary normal vector on E with respect to K.
This definition serves to guarantee opposite sign values for $\epsilon_{K,E}$ and $\epsilon_{K',E}$, i.e. $\epsilon_{K,E} = -\epsilon_{K',E} = \pm 1$, $\forall\, E = K \cap K'$. Thus, continuities of pressure and flux are satisfied by imposing for every $E = K \cap K'$ the following

$$q_E = \epsilon_{K,E} q_{K,E} = \epsilon_{K',E} q_{K',E}, \tag{22}$$

$$\epsilon_{K,E} tp_{K,E} + \epsilon_{K',E} tp_{K',E} = 0. \tag{23}$$

By multiplying the algebraic equations (8) by $\epsilon_{K,E}$, we get

$$\epsilon_{K,E}\, tp_{K,E} = \epsilon_{K,E}\, p_K - \sum_{E' \subset \partial K} (B_K)^{\epsilon}_{E,E'}\, q_{E'} \qquad E \subset \partial K, \tag{24}$$

where $(B_K)^{\epsilon}_{E,E'} = \epsilon_{K,E}(B_K)_{E,E'}\, \epsilon_{K,E'}$ and $q_{E'} = \epsilon_{K,E'} q_{K,E'}$.
Therefrom, it is possible to eliminate the unknowns $tp_{K,E}$ by plugging (24) into (23). By taking into account the boundary conditions, the algebraic system with unknowns the pressure vector P and the flux vector Q can be written in the matrix form

$$\tilde{R}^T\, P - \tilde{M}\, Q - \tilde{V} = 0, \tag{25}$$

where

- $\tilde{R}^T$ and $\tilde{M}$ are two matrices whose structures are similar to those defined in (12). $\forall E, E' \in \mathcal{E}_h$, $K \in \mathcal{T}_h$, their entries are given by

$$(\tilde{R})_{K,E} = \epsilon_{K,E}, \qquad (\tilde{M})_{E,E'} = \sum_{\partial K \supset E,E'} (B_K)^{\epsilon}_{E,E'}\,;$$

- $\tilde{V}$ is a vector corresponding to the Dirichlet and Neumann boundary conditions.

For the discretized mass conservation equation (13), it can be written in the matrix form as follows

$$S\, \frac{dP}{dt} + \tilde{R}\, Q = \tilde{F}, \tag{26}$$

where S is the same matrix as that defined in (14), and $\tilde{F}$ corresponds to the source/sink function as well as to the imposed fluxes given by the Neumann boundary conditions.

2.3.1 Properties of the algebraic systems obtained from the MFEM

The two systems (26) and (25) can be globally written in the matrix form

$$\mathcal{L} \begin{pmatrix} \frac{dP}{dt} \\ \frac{dQ}{dt} \end{pmatrix} - \tilde{\mathcal{J}} \begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \tilde{F} \\ \tilde{V} \end{pmatrix}, \tag{27}$$

where $\tilde{\mathcal{J}} = \begin{pmatrix} 0 & -\tilde{R} \\ -\tilde{R}^T & \tilde{M} \end{pmatrix}$.
In a similar process as in the MHFE, we use backward Euler scheme for the temporal

discretization of (27) to get

$$\left(\mathcal{L} - \Delta t\, \tilde{\mathcal{J}}\right) \begin{pmatrix} P^n \\ Q^n \end{pmatrix} = \mathcal{L} \begin{pmatrix} P^{n-1} \\ Q^{n-1} \end{pmatrix} + \Delta t \begin{pmatrix} \tilde{F}^n \\ \tilde{V}^n \end{pmatrix}. \tag{28}$$

It should be noted that the choice of a numerical solver of the above system is restrained by the fact that its coefficient matrix $(\mathcal{L}-\Delta t\, \tilde{\mathcal{J}})$ is symmetric but indefinite. Since the diagonal matrix S is invertible, it is easy to separate the unknowns P and Q by constructing the Schur complement system, i.e.

$$\begin{cases} (\tilde{M} + \Delta t\, \tilde{R}^T S^{-1} \tilde{R}) Q^n = \tilde{R}^T (P^{n-1} + \Delta t\, S^{-1}\, \tilde{F}^n) + \tilde{V}^n. \\ S\, P^n = S\, P^{n-1} - \Delta t\, \tilde{R}\, Q^n + \Delta t\, \tilde{F}^n. \end{cases} \tag{29}$$

Lemma 2.3 *The matrix B_K^ϵ is positive definite.*

Proof:

Let $x \in \mathbb{R}^3$ be a nonzero vector, then

$$x^T B_K^\epsilon\, x = \sum_{E,E' \subset \partial K} x_E \epsilon_{K,E} (B_K)_{E,E'}\, \epsilon_{K,E'}\, x_{E'} = \tilde{x}^T B_K\, \tilde{x} > 0,$$

where $\tilde{x} = (x_E \epsilon_{K,E})_{E \subset \partial K}$.
Further, one can easily show that $\tilde{e}_K = (\epsilon_{K,E})_{E \subset \partial K}$ is an eigenvector of B_K^ϵ and $\frac{3}{\alpha_K}$ is the corresponding eigenvalue.

□

Lemma 2.4 *The matrix $\tilde{M}$ is positive definite.*

Proof:

Similar to lemma (2.2).

□

Proposition 2.3 *The Schur complement matrix $(\tilde{M} + \Delta t\, \tilde{R}^T S^{-1} \tilde{R})$ is positive definite.*

Proof:

Since $\tilde{M}$ is positive definite and $\Delta t(\tilde{R}^T S^{-1} \tilde{R})$ is semidefinite then the Schur complement matrix is positive definite.

□

As a result, the problem is reduced to solve at each time step two linear systems of similar properties to those derived by the MHFE formulation.
The principal steps of the MFE algorithm are illustrated in (Algo. 2).

Algorithm 2 *Principal steps of the MFE algorithm.*

1– Initialize geometry and physical parameters of the problem.
2– Create the Schur complement matrix.
3– Iterations on the time-steps.
4– Find Q by solving the first system in (29).
5– Find P by solving the second system in (29).
6– Write the outputs P and Q.

As a matter of fact, to avoid inverting the elementary matrix B_K is an important advantage of the MFE over the MHFE from a computational point of view.

2.3.2 Discretization of the time independent problem

In a similar manner, the algebraic system of the time independent problem (2) derived from the MFE approximation is given by

$$\begin{pmatrix} 0 & -\tilde{R} \\ -\tilde{R}^T & \tilde{M} \end{pmatrix} \begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \tilde{F} \\ \tilde{V} \end{pmatrix}. \tag{30}$$

This linear system is relatively large compared to (19), furthermore its resolution is restricted by the fact that the coefficient matrix is symmetric but indefinite.

3 Effects of the mesh geometry on the mixed and mixed-hybrid approximated solutions

The gradual evolution of symbolic programming languages has enriched not only the technical computing but also the numerical computations in solving various problems in applied sciences. However, many hindrances are still restricting this programming technique. Commonly, symbolic computation programs are regarded as fairly limited in solving PDEs where analytical solutions may not exist and therefore may not be implemented in the symbolic computation packages. Furthermore, numerical computations in symbolic languages (being interpreted languages) are limited in use for large scale problems, besides by comparing with compiled languages, symbolic

computing usually involves high computational overhead. Subsequently, one of the used approaches to get benefits of technical languages is to translate the symbolic code into a more efficient computational environment and this can also be done automatically. Thus, the outfitted translated routine can be thrown in the compiled language (e.g. C or Fortran) and even can be treated as a black-box. Although the theoretical results of the algorithm are correct, one cannot guarantee the accuracy of the numerical results since the algorithm may fail to preserve its numerical stability.

This section is devoted to inspect the numerical reliability of two matrix-inversion subroutines used to compute and invert the elementary matrices $B_{_K}$ defined in (10). The first one, which was originally utilized in our Fortran 77 MHFE program, was generated automatically by using mainly the two Maple functions *inverse* and *fortran*. The matrix-inversion function of Maple is based on Cramer's rule which is well known to be unstable (see [8]). Accordingly, we have detected many numerical examples where non-consistent results obtained by the MHFE code are mainly caused by this subroutine. On the other hand, our reconstructed subroutine is based on the LDL^T–factorization method.
Being $B_{_K}$ a symmetric matrix, it can be decomposed into the form $B_{_K} = LDL^{^T}$ where L is a lower triangular matrix with unitary diagonal and D is a diagonal matrix. Consequently, the inverted matrix $B_{_K}^{-1}$ can be easily computed by solving respectively lower, diagonal and upper linear systems of the form

$$B_{_K}^{-1} = (L^{-T} D^{-1} L^{-1}).$$

In this case, the positive definiteness of $B_{_K}$ ensures the numerical stability of this method (see, e.g., [7]). Nevertheless, the total number of arithmetic operations is reduced by this method to about 110 after it was about 230 arithmetic operations in the former method. Accordingly, it is found that the LDL-subroutine is at least two times faster than Cramer-subroutine.
Despite the fact that the convergence of the MFE/MHFE method does not necessitate the Delaunay triangulation conditions [15], we give some numerical examples showing that the MHFE method suffers with the presence of badly shaped discretized elements. In contrast, the MFE method does not face this numerical difficulty.
In the numerical example depicted in Fig.2, the domain is discretized into a nonuniform mesh of triangular elements. Throughout the incoming tests the domain is taken to be homogeneous with unit conductivity tensor and storage coefficient and without sink/source terms, the boundary conditions are time-independent constant functions (see Fig.2), and the simulation sampling time interval is $]0, 3]$. These numerical tests intend to study the consequences in the case of bad mesh quality, i.e.

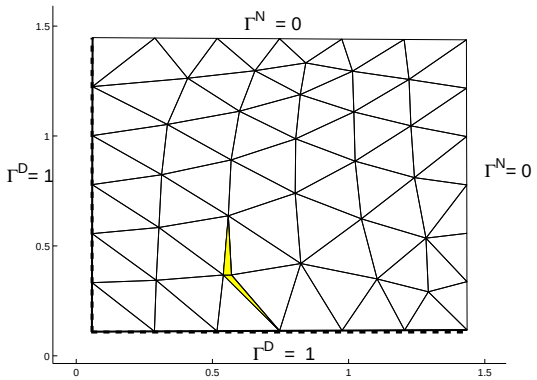

Figure 2: Nonuniform mesh with two flat triangular elements,
with $T = 3,\ \Delta t = 1,\ s = 1,\ \mathcal{K} = I_{2\times 2},\ f = 0.$

meshes containing badly shaped elements which are often obtained by automatic or adaptive refinements. According to the definition given in [17], the quality of a triangular element K is evaluated by

$$\mathcal{Q}(K) = \theta\ \frac{\rho_{_K}}{h_{_K}}, \tag{31}$$

where $h_{_K}$ is the element diameter (the length of its longest edge), $\rho_{_K}$ is the in-radius and θ is a scaling factor such that the quality of an equilateral triangle is 1 ($\theta = 2\sqrt{3}$).
Thus, a triangle quality varies in the interval $]0, 1]$, at worst close to 0, at best equal to 1. The mesh quality is measured by the qualities of its worst elements and their geometrical distribution. The two shaded triangles appearing in Fig.2 are artificially constructed so that we can increase their flatness by decreasing their common edge.

In Fig.3, we compare the computed results obtained by the MHFE code where the two matrix-inversion subroutines are used, the first (Fig.3a) is the one generated automatically by the symbolic language and the second (Fig.3b) is based on the LDL–factorization method. For a mesh quality about 0.2, the velocity flow and pressure contour obtained by both methods are apparently undifferentiated. However, by lessening the mesh quality to $\approx 10^{-5}$ in Fig.4, we can clearly notice that, unlike the LDL–method (Fig.4b), the other code (Fig.4a) brings on senseless results.

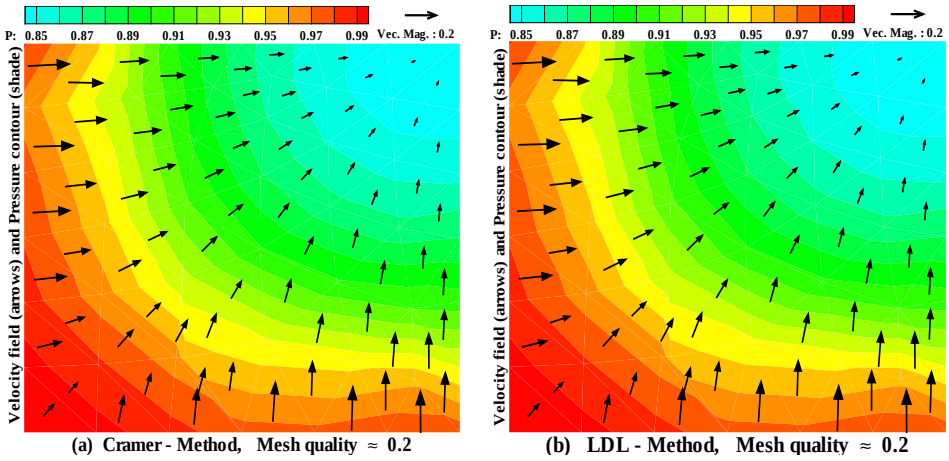

Figure 3: MHFE numerical results obtained by using Cramer and LDL–subroutines.

The cause of this shortcoming is the instability of Cramer-method where mistaken results can be obtained especially when inverting ill-conditioned matrices.

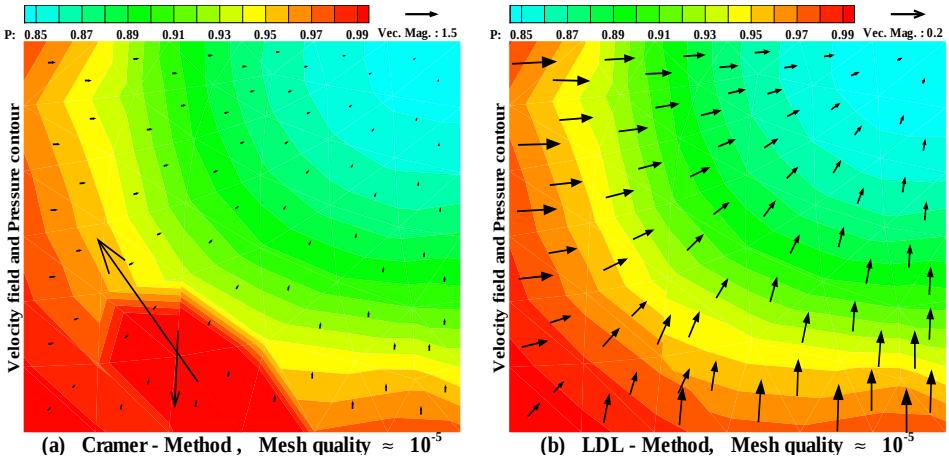

Figure 4: MHFE numerical results obtained by using Cramer and LDL–subroutines.

3.1 Numerical stability of Cramer and LDL^T matrix-inversion subroutines

This section is devoted to check out the stability of the two subroutines by using Aquarels which is a software toolbox used to evaluate the reliability of numerical algorithms [18, 19]. One of the tools available in Aquarels depends on functional stability analysis so that the instability of a given program is automatically detected by using perturbation techniques. In this case, the code is treated as a black-box, some directives are added to declare the input parameters to be perturbed and the output variables to be analyzed. This approach has the practical advantage of being easy to use and flexible to test general numerical algorithms. The amplitude of perturbations is specified by indicating the number and the position of bits in the variable mantissa to be changed. Thus, by executing the code several times according to a specified number of samples, the code stability can be estimated by examining the problem conditioning as well as the forward errors. These perturbed samples of the code are automatically generated and analyzed by Aquarels.
In order to eschew from any numerical interventions, the two subroutines are investigated individually and independently of their main code. The stability analysis of Aquarels is based on *perturbing* the inputs which are the three vertices of a triangle and *analyzing* the output which is the inverted matrix. The graphical results depicted in Fig.5 are automatically generated by Aquarels and the numerical errors are tabulated in table 1. The number of perturbation samples is chosen to be 16. In practice, a numerical algorithm is expected to be stable for a given problem if linear interpolated line is obtained, i.e. slope (Regularité) is close to one (see Fig.5). In test **1**, almost an equilateral triangle is taken (triangle quality ≈ 0.9). In this case, both routines are stable and even they sound equivalent (Regularité ≈ 1). On the other hand, the instability of Cramer-subroutine is clearly detected by increasing the flatness of triangles in tests **2** and **3**, beyond the other subroutine is always stable.

Table 1: Absolute and relative errors of the two methods measured by using $\|.\|_\infty$.

	Absolute Error		Relative Error	
	Cramer–Method	LDL–Method	Cramer–Method	LDL–Method
Test **1**	1.7×10^{-15}	4.4×10^{-16}	5.9×10^{-16}	1.5×10^{-16}
Test **2**	$1.6 \times 10^{+03}$	4.7×10^{-03}	1.1×10^{-04}	3.3×10^{-10}
Test **3**	$3.5 \times 10^{+07}$	2.7×10^{-02}	1.6×10^{-01}	1.2×10^{-10}

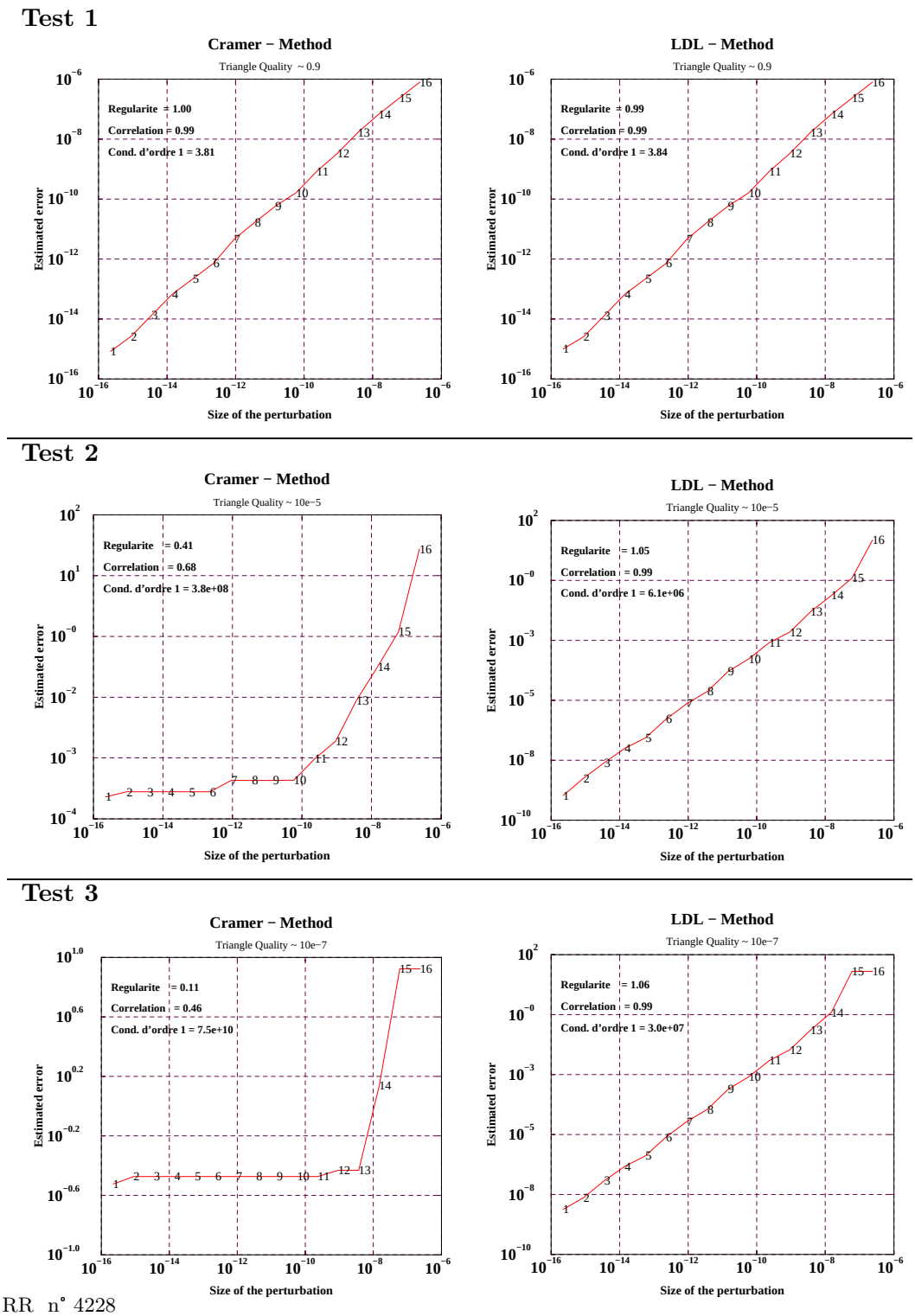

Figure 5: Different stability tests analyzed by Aquarels.

In table 1, the absolute and relative errors of the three numerical experiments are estimated by comparing the subroutines' solutions with the solutions computed formally with very high-precision arithmetic. In is clear that the numerical errors caused by the instability of Cramer's rule could indeed afflict the output of the main algorithm.

Despite the fact that the subroutine based on the LDL–method is stable, unfortunately, big direct errors may also be obtained since ill-conditioned problems could build on the error and consequently no numerical algorithm with fixed precision floating point computation is able to guarantee very accurate results. In Fig.6a, the MHFE code leads to inconsistent results with the use of LDL–subroutine. In this example, the triangle quality is about 10^{-8} and the condition number of the matrix to invert is about 10^{15} which is nearby the limits of the machine precision. Nevertheless, the aim of presenting this example is to show that even with such elements of very bad quality which afflict the MHFE algorithm, this does not cause any shortcoming of the accuracy of results obtained by the MFE code (see Fig.6b) simply because the inversion of the elementary matrices is needless.

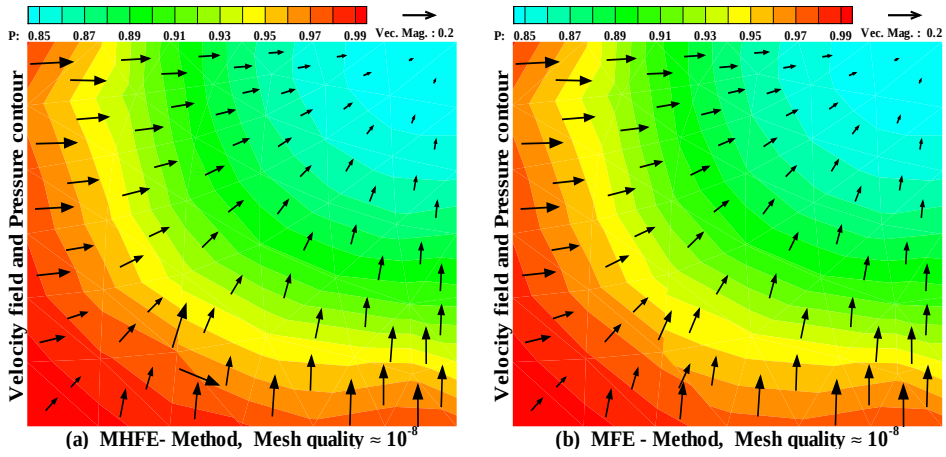

Figure 6: Numerical results computed by using MHFE and MFE codes.

4 Functioning of MFE and MHFE methods on heterogeneous media

In the previous section, we have investigated the numerical difficulties induced by the mesh geometry. This section is devoted to the numerical behavior of the approximated solutions of the mixed methods in heterogeneous media. We are interested in the case of rough physical parameters, specifically physical problems to which large jumps in the tensor of conductivity or small values of the storage coefficients are imposed. It is found that such problems affect the numerical accuracy and time requirements of the MFE and MHFE algorithms, differently. The first common numerical difficulty, which is caused by sharp leaps of the conductivity tensor, is the growth of the conditioning of the algebraic linear systems to solve.

Suppose henceforth that the heterogeneous domain Ω is composed of a set of sub-domains Ω_i, $i = 1, \ldots, n$, according to their conductivity tensors $\mathcal{K}_i$, i.e. each Ω_i has a homogeneous conductivity (see Fig.7). We assume that Ω is uniformly

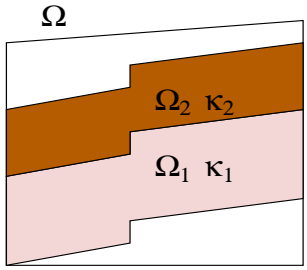

Figure 7: Decomposition of the domain according to the permeability values.

discretized and the medium is isotropic, i.e. the tensor of conductivity is equal to a scalar function times the unit tensor. By κ_i, we denote the scalars such that $\mathcal{K}_i = \kappa_i \, I_{2\times 2}$. We also set

$$\frac{\kappa_2}{\kappa_1} = \max \left\{ \frac{\kappa_i}{\kappa_j} \mid \Omega_i \cap \Omega_j \neq \emptyset, \; i \neq j, \; i, j = 1, \ldots, n \right\}.$$

For the sake of simplicity, we suppose that $\Omega = \Omega_1 \cup \Omega_2$. The aim here is to give some estimations of the conditioning of the algebraic systems induced by both mixed methods.

First, let us prove the following lemmas.

Lemma 4.1 *Let A (resp. B) be a non-singular (resp. singular) matrix, then the condition number $\mathcal{X}(A)$ of A is bounded by the following inequality*

$$\mathcal{X}(A) \geq \frac{\|A\|}{\|A-B\|}.$$

Proof:

Since B is a singular matrix, there exists a nonzero vector x such that $Bx = 0$, then we can write

$$\|Ax\| = \|(A-B)x\| \leq \|A-B\|\|x\|, \quad (32)$$
$$\|x\| = \|A^{-1}Ax\| \leq \|A^{-1}\|\|Ax\|, \quad (33)$$

Hence, the lemma is a direct conclusion of (32) and (33) .

□

In the runs, we'll use the norms $\|.\|_1$ or $\|.\|_\infty$ and assume that the fraction $\frac{\kappa_2}{\kappa_1} \gg 1$.

4.1 The condition number estimations of the derived MHFE algebraic systems

Lemma 4.2 *Let $\mathcal{J}_{\kappa_1,\kappa_2}$ denote the matrix $\mathcal{J}$ in the algebraic system (16) fitted out by the MHFE formulation over the heterogeneous domain Ω (Fig.7). Similarly, we use the notation $\mathcal{J}_{\kappa_i}$ if the medium is homogeneous with hydraulic conductivity κ_i, i.e $\Omega_i = \Omega$. Then, we get*

$$\|J_{\kappa_1}\| \quad is \quad \mathcal{O}(\kappa_1),$$
$$\|J_{\kappa_1,\kappa_2}\| \quad is \quad \mathcal{O}(\kappa_2).$$

Proof:

By expressing explicitly the hydraulic conductivity parameter $\kappa_{_K}$ in the definition of $B_{_K}$ given in (10), we get

$$B_{_K} = \frac{1}{\kappa_{_K}} \widehat{B}_{_K}, \quad (34)$$

such that, $\widehat{B}_K$ is independent of κ_K.
In a similar manner, $\forall K,\ K' \in \mathcal{T}_h,\ E \subset \partial K,\ E' \subset \partial K'$, we can write

$$\begin{array}{ll} R_{K,E} & = \kappa_K\, \widehat{\alpha}_{K,E}, \\ D_K & = \kappa_K\, \widehat{\alpha}_K, \\ M_{E,E'} & = \left\{ \begin{array}{ll} \kappa_K\, (\widehat{B}_K^{-1})_{E,E'} & E \neq E', \\ \kappa_K [(\widehat{B}_K^{-1})_{E,E} + \frac{\kappa_{K'}}{\kappa_K}\, (\widehat{B}_{K'}^{-1})_{E,E}] & E = E'. \end{array} \right. \end{array} \tag{35}$$

Then, in the homogeneous medium case, it is evident that $\|R_{\kappa_1}\|$, $\|D_{\kappa_1}\|$ and $\|M_{\kappa_1}\|$ are $\mathcal{O}(\kappa_1)$ and so is $\|\mathcal{J}_{\kappa_1}\|$. Similarly, in the heterogeneous medium case, by using ∞-norm (or 1-norm), one can easily verify that $\|R_{\kappa_1,\kappa_2}\|$, $\|D_{\kappa_1,\kappa_2}\|$, $\|M_{\kappa_1,\kappa_2}\|$ and $\|\mathcal{J}_{\kappa_1,\kappa_2}\|$ are $\mathcal{O}(\kappa_2)$.

□

Proposition 4.1 *The condition numbers of the algebraic systems induced by the MHFE method have the following bounds:*

1. $\mathcal{X}(\mathcal{J}_{\kappa_1,\kappa_2}) \geq \frac{\|\mathcal{J}_{\kappa_1,\kappa_2}\|}{\|\mathcal{J}_{\kappa_1}\|}$, *which is* $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$.
2. $\mathcal{X}(\mathcal{L} - \Delta t\, \mathcal{J}_{\kappa_1,\kappa_2})$ *is* $\mathcal{O}\left(\max\left\{\frac{1}{\Delta t}\, \frac{\|S\|}{\kappa_1},\ \frac{\kappa_2}{\kappa_1}\right\}\right)$.
3. $\mathcal{X}(\mathcal{A}_{\kappa_1,\kappa_2})$ *is* $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$, *where* $\mathcal{A}_{\kappa_1,\kappa_2} = (M - \Delta t\, R^T G^{-1} R)_{\kappa_1,\kappa_2}$ *is the Schur complement matrix for the time dependent problem given in (18).*
4. $\mathcal{X}(\mathcal{B}_{\kappa_1,\kappa_2})$ *is* $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$, *where* $\mathcal{B}_{\kappa_1,\kappa_2} = (M - R^T D^{-1} R)_{\kappa_1,\kappa_2}$ *is the Schur complement matrix for the time independent problem given in (20).*

Proof:

Proof of 1. Referring to (35), it is evident to see that for $i = 1, 2$, we have

$$\begin{array}{ll} R_{\kappa_1,\kappa_2} & = R_{\kappa_i} \quad \text{in } \Omega_i, \\ D_{\kappa_1,\kappa_2} & = D_{\kappa_i} \quad \text{in } \Omega_i, \\ M_{\kappa_1,\kappa_2} & = \left\{ \begin{array}{ll} M_{\kappa_i} & \text{in } \Omega_i, \\ M_{\kappa_1,\kappa_2} & \text{on } \Omega_1 \cap \Omega_2. \end{array} \right. \end{array}$$

Accordingly, $\mathcal{J}_{\kappa_1,\kappa_2}$ can be expressed as

$$\mathcal{J}_{\kappa_1,\kappa_2} = \begin{cases} \mathcal{J}_{\kappa_i} & \text{in } \Omega_i, \;\; i = 1, 2, \\ \mathcal{J}_{\kappa_1,\kappa_2} & \text{on } \Omega_1 \cap \Omega_2. \end{cases}$$

Now, let's define the following matrix

$$J_{\kappa_1,\kappa_2} = \begin{cases} 0 & \text{in } \Omega_1, \\ \mathcal{J}_{\kappa_2} - \mathcal{J}_{\kappa_1} & \text{in } \Omega_2, \\ \mathcal{J}_{\kappa_1,\kappa_2} - \mathcal{J}_{\kappa_1} & \text{on } \Omega_1 \cap \Omega_2. \end{cases}$$

This matrix is constructed in such a way so that $\mathcal{J}_{\kappa_1,\kappa_2} - J_{\kappa_1,\kappa_2} = \mathcal{J}_{\kappa_1}$. To prove the singularity of J_{κ_1,κ_2}, it is sufficient to find one edge in Ω_1 not belonging to Ω_2, then the corresponding row in J_{κ_1,κ_2} is null. Hence, the sought inequality is a direct consequence of lemma 4.1. Furthermore, by applying lemma 4.2, the condition number $\mathcal{X}(\mathcal{J}_{\kappa_1,\kappa_2})$ is $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$.

Proof of 2. Since $\|\mathcal{J}_{\kappa_1,\kappa_2}\| = \mathcal{O}(\kappa_2)$, there exists $C \in \mathbb{R}$ such that $\|\mathcal{J}_{\kappa_1,\kappa_2}\| = C\,\kappa_2$. It follows that

$$\begin{aligned} \|\mathcal{L} - \Delta t\,\mathcal{J}_{\kappa_1,\kappa_2}\| &\leq \|\mathcal{L}\| + \Delta t\,\|\mathcal{J}_{\kappa_1,\kappa_2}\| \\ &\leq \|S\| + C\,\Delta t\,\kappa_2 \\ &\leq 2\max\{\|S\|,\, C\,\Delta t\,\kappa_2\}. \end{aligned}$$

Thus, $\|\mathcal{L} - \Delta t\,\mathcal{J}_{\kappa_1,\kappa_2}\|$ is $\mathcal{O}\left(\max\{\|S\|,\, \Delta t\,\kappa_2\}\right)$.
On the other hand, we can easily show that

$$(\mathcal{L} - \Delta t\,\mathcal{J}_{\kappa_1,\kappa_2}) + \Delta t\,\mathcal{J}_{\kappa_1} = \mathcal{L} - \Delta t(\mathcal{J}_{\kappa_1,\kappa_2} - \mathcal{J}_{\kappa_1})$$

is a singular matrix (similar to the previous proof). Therefore, the sought inequality is a direct conclusion of lemma (4.1), i.e.

$$\mathcal{X}(\mathcal{L} - \Delta t\,\mathcal{J}_{\kappa_1,\kappa_2}) \geq \frac{\|(\mathcal{L} - \Delta t\,\mathcal{J}_{\kappa_1,\kappa_2})\|}{\Delta t\,\|\mathcal{J}_{\kappa_1}\|} = \mathcal{O}\left(\frac{\max\left\{\frac{1}{\Delta t}\,\|S\|,\, \kappa_2\right\}}{\kappa_1}\right).$$

Proof of 3. $\forall K \in \mathcal{T}_h$, $E, E' \subset \partial K$, we have

$$\begin{aligned} |\,(\mathcal{A}_{\kappa_1})_{E,E'}\,| &= |\,\kappa_1\,\widehat{M}_{E,E'} - \sum_{\partial K \supset E,E'} \left(\frac{\Delta t\,\kappa_1}{s_K + \Delta t\,\kappa_1}\widehat{R}_{K,E'}\,\widehat{R}_{K,E}\right)| \\ &\leq \kappa_1(|\,\widehat{M}_{E,E'}\,| + \sum_{\partial K \supset E,E'} \frac{1}{\frac{s_K}{\Delta t\,\kappa_1} + 1}\;|\,\widehat{R}_{K,E'}\,\widehat{R}_{K,E}\,|) \\ &\leq \kappa_1(|\,\widehat{M}_{E,E'}\,| + \sum_{\partial K \supset E,E'} |\,\widehat{R}_{K,E'}\,\widehat{R}_{K,E}\,|). \end{aligned}$$

Thus, $\|\mathcal{A}_{\kappa_1}\|$ is $\mathcal{O}(\kappa_1)$. In the heterogeneous case, we can also verify that $\|\mathcal{A}_{\kappa_1,\kappa_2}\|$ is $\mathcal{O}(\kappa_2)$. Hence, our proposition holds by showing that $(\mathcal{A}_{\kappa_1,\kappa_2} - \mathcal{A}_{\kappa_1})$ is a singular matrix then applying lemma 4.1.

Proof of 4. Similar to the previous proof.

□

4.2 The condition number estimations of the derived MFE algebraic systems

Proposition 4.2 *The condition numbers of the algebraic systems induced by the MFE method have the following bounds:*

1. $\mathcal{X}(\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}) \geq \frac{\|\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|}{\|\tilde{\mathcal{J}}_{\kappa_2}\|}$, *which is* $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$.
2. $\mathcal{X}(\mathcal{L} - \Delta t\, \tilde{\mathcal{J}}_{\kappa_1,\kappa_2})$ *is* $\mathcal{O}\left(\max\left\{\frac{\kappa_2}{\Delta t}\,\|S\|, \frac{\kappa_2}{\kappa_1}\right\}\right)$.
3. $\mathcal{X}(\tilde{\mathcal{A}}_{\kappa_1,\kappa_2})$ *is* $\mathcal{O}\left(\max\left\{\kappa_1\,\Delta t\,\|S^{-1}\|,\ \frac{\kappa_2}{\kappa_1}(\frac{1+\kappa_1\,\Delta t\,\|S^{-1}\|}{1+\kappa_2\,\Delta t\,\|S^{-1}\|})\right\}\right)$, *where* $\tilde{\mathcal{A}}_{\kappa_1,\kappa_2} = (\tilde{M}_{\kappa_1,\kappa_2} + \Delta t\,\tilde{R}^T S^{-1}\tilde{R})$ *is the Schur complement matrix for the time dependent problem given in (29).*

Proof:

Proof of 1. In a similar manner to (34) and (35), we can write

$$B^{\epsilon}_{K} = \frac{1}{\kappa_K}\,\widehat{B}^{\epsilon}_{K},\quad \tilde{M}_{E,E'} = \begin{cases} \frac{1}{\kappa_K}\,(\widehat{B}^{\epsilon}_{K})_{E,E'} & E \neq E', \\ \frac{1}{\kappa_K}((\widehat{B}^{\epsilon}_{K})_{E,E'} + \frac{\kappa_K}{\kappa_{K'}}\,(\widehat{B}^{\epsilon}_{K'})_{E,E}) & E = E', \end{cases}$$

such that, $\widehat{B}^{\epsilon}_{K}$ is independent of κ_K.
Thus, one can easily verify that $\|\tilde{\mathcal{J}}_{\kappa_2}\|$ and $\|\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|$ are $\mathcal{O}(\frac{1}{\kappa_2})$ and $\mathcal{O}(\frac{1}{\kappa_1})$, respectively. Moreover, the matrix

$$\tilde{\mathcal{J}}_{\kappa_1,\kappa_2} - \tilde{\mathcal{J}}_{\kappa_2} = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{M}_{\kappa_1,\kappa_2} - \tilde{M}_{\kappa_2} \end{pmatrix}$$

is singular. Therefore, by applying lemma 4.1, we get

$$\mathcal{X}(\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}) \geq \frac{\|\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|}{\|\tilde{\mathcal{J}}_{\kappa_2}\|} = \mathcal{O}(\frac{\kappa_2}{\kappa_1}).$$

Proof of 2. We have

$$\begin{aligned}\|\mathcal{L}-\Delta t\,\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\| &\leq \|\mathcal{L}\|+\Delta t\,\|\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|\\ &\leq \|S\|+\Delta t\,\frac{C}{\kappa_1}\\ &\leq 2\max\left\{\|S\|,\ \Delta t\,\frac{C}{\kappa_1}\right\},\ \text{for some } C\in\mathbb{R}.\end{aligned}$$

Consequently, $\|\mathcal{L}-\Delta t\,\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|$ is $\mathcal{O}\left(\max\left\{\|S\|,\ \frac{\Delta t}{\kappa_1}\right\}\right)$. On the other hand, the matrix $(\mathcal{L}-\Delta t\,\tilde{\mathcal{J}}_{\kappa_1,\kappa_2})+\Delta t\,\tilde{\mathcal{J}}_{\kappa_2}$ is singular, then by applying lemma 4.1, we get

$$\mathcal{X}(\mathcal{L}-\Delta t\,\tilde{\mathcal{J}}_{\kappa_1,\kappa_2})\geq\frac{\|\mathcal{L}-\Delta t\,\tilde{\mathcal{J}}_{\kappa_1,\kappa_2}\|}{\|\Delta t\,\tilde{\mathcal{J}}_{\kappa_2}\|}=\mathcal{O}\left(\max\left\{\frac{\kappa_2}{\Delta t}\,\|S\|,\frac{\kappa_2}{\kappa_1}\right\}\right).$$

Proof of **3.** We have

$$\|\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}\| \geq \|\tilde{M}_{\kappa_1,\kappa_2}\|+\Delta t\,\|\tilde{R}^T S^{-1}\tilde{R}\|=\mathcal{O}(\frac{1}{\kappa_1}+\Delta t\,\|S^{-1}\|).$$

Since the matrix $\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}-\tilde{M}_{\kappa_1,\kappa_2}=\Delta t\,\tilde{R}^T S^{-1}\tilde{R}$ is rank deficient (not of full rank), so it is singular. Thus, by applying lemma 4.1, we get

$$\begin{aligned}\mathcal{X}(\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}) &\geq \frac{\|\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}\|}{\|\tilde{M}_{\kappa_1,\kappa_2}\|}\\ &= \frac{\mathcal{O}\left(\frac{1}{\kappa_1}+\Delta t\,\|S^{-1}\|\right)}{\mathcal{O}\left(\frac{1}{\kappa_1}\right)}=\mathcal{O}\left(\kappa_1\,\Delta t\,\|S^{-1}\|\right).\end{aligned}\tag{36}$$

On the other hand, the matrix $\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}-\tilde{\mathcal{A}}_{\kappa_2}=(\tilde{M}_{\kappa_1,\kappa_2}-\tilde{M}_{\kappa_2})$ is singular, where $\tilde{\mathcal{A}}_{\kappa_2}=\tilde{M}_{\kappa_2}+\Delta t\,\tilde{R}^T S^{-1}\tilde{R}$). Thus, we also get

$$\begin{aligned}\mathcal{X}(\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}) &\geq \frac{\|\tilde{\mathcal{A}}_{\kappa_1,\kappa_2}\|}{\|\tilde{\mathcal{A}}_{\kappa_2}\|}\\ &= \frac{\mathcal{O}\left(\frac{1}{\kappa_1}+\Delta t\,\|S^{-1}\|\right)}{\mathcal{O}\left(\frac{1}{\kappa_2}+\Delta t\,\|S^{-1}\|\right)}=\mathcal{O}\left(\frac{\kappa_2}{\kappa_1}\,\frac{1+\kappa_1\,\Delta t\,\|S^{-1}\|}{1+\kappa_2\,\Delta t\,\|S^{-1}\|}\right).\end{aligned}\tag{37}$$

Therefore, the sought inequality is a conclusion of (36) and (37).

□

Remark 4.1 *Proposition 4.1 indicates that the value of the storage coefficient s has a critical effect on the solution of the linear system (29) associated to the MFE formulation. As s tends to zero, the condition number of the coefficient matrix in (29) blows up to infinity. In this case, it is preferred to solve the symmetric indefinite linear system (28). While on the contrary, the MHFE formulation does not face this difficulty.*

Remark 4.2 *Suppose that the mesh is made up of right angle triangles that are constructed by subdivisions of rectangular elements, then in (34), we can bring out a term which depends on the geometry of the elements (see [14]), i.e. (34) can be expressed as*

$$B_{_K} = \frac{1}{\kappa_{_K}} \frac{\Delta x_{_K}}{\Delta y_{_K}} \widehat{B}_{_K},$$

such that, $\widehat{B}_{_K}$ *is independent of the element geometry.*
Hence, one can notice that the ratio $\frac{\Delta x_{_K}}{\Delta y_{_K}}$*, which is related to the element quality, plays a similar role as the conductivity parameter* $\kappa_{_K}$*. In other words, large variations among the qualities of the elements enlarge the conditioning as well.*

4.3 Accumulation of numerical errors by the MHFE formulation

The mixed finite element methods have gained a big popularity chiefly for two superior advantages. First, the two physical quantities, the pressure and the flux, are computed with very accurate approximations and with the same order of convergence. Second, they conserve mass locally. Unfortunately, numerical experiments showed that the accumulation of numerical errors could break down the theory. In this section, we address another computational difficulty which could afflict the accuracy of the computed flux as well as the local mass balance property. This numerical problem concerns specifically the MHFE algorithm and it is mostly caused by large jumps in the permeability parameters. After computing the pressure and its traces, the flux, in the MHFE formulation, is computed by the local equation (9). By rewriting this equation $\forall K \in \mathcal{T}_h$, $E \subset \partial K$, we get

$$\begin{aligned} q_{K,E} &= \kappa_{_K} \left[\widehat{\alpha}_{K,E}\, p_{_K} - \sum_{E' \subset \partial K} (\widehat{B}_{_K}^{-1})_{E,E'}\, tp_{K,E'} + \xi_{K,E} \right] \\ &= \bar{q}_{K,E} + \kappa_{_K}\, \xi_{K,E}, \end{aligned} \tag{38}$$

where $\bar{q}_{K_i,E}$ is supposed to be the exact value of the flux across E, and $\xi_{K,E}$ denotes the numerical errors accumulated while computing $q_{K,E}$, i.e. it is the roundoff error plus the truncation error coming out from solving the linear systems. Thus, the sum of fluxes through the edges of K becomes

$$\sum_{E\subset\partial K} q_{K,E} = \sum_{E\subset\partial K} \left[\bar{q}_{K,E} + \kappa_K \xi_{K,E}\right]. \tag{39}$$

As the simulation time increases, the transient solution converges toward the stationary one. Theoretically, the local mass conservation property necessitates that the sum of fluxes over each element be equal to the imposed local sink/source term (13). However, with the presence of numerical errors, this can be expressed as follows

$$\left[\sum_{E\subset\partial K} q_{K,E} - f_K\right] = \kappa_K \sum_{E\subset\partial K} \xi_{K,E}. \tag{40}$$

Consequently, the numerical difficulty in computing (38) and then (40) is twofold.

1. The hoped for computed results may not be so reliable since multiplications by rough conductivity parameters could afflict the approximated fluxes.
2. The flux over the grid elements is not calculated with the same order of accuracy. This can be expressed in the following proposition.

Proposition 4.3 *The condition number of the algebraic system whereby the flux is computed is* $\mathcal{O}\left((\frac{\kappa_2}{\kappa_1})^2\right)$.

Proof:

By inverting B_K in (9), we obtain

$$Q_K = \kappa_K \, \widehat{B}_K^{-1}(p_K \, e - T_{P_K}), \qquad K \in \mathcal{T}_h. \tag{41}$$

These equations can be gathered in the matrix form

$$Q = B\,(\tilde{P} - \tilde{T_P}), \tag{42}$$

where
$\tilde{P}$ and $\tilde{T_p}$ are $3\mathcal{N}_{\mathcal{T}}$–dimensional vectors, such that

$$\tilde{P} = (p_K \, e)_{K\in\mathcal{T}_K}, \;\; \tilde{T_p} = (T_{P_K})_{K\in\mathcal{T}_K};$$

B is a $3\mathcal{N}_{\mathcal{T}} \times 3\mathcal{N}_{\mathcal{T}}$ block-diagonal matrix, such that each block corresponds to an element $K \in \mathcal{T}_K$ and is equal to $(\kappa_K \widehat{B}_K^{-1})$.
By using lemma 4.1, one can easily verify that $\mathcal{X}(B) = \mathcal{O}(\frac{\kappa_2}{\kappa_1})$. On the other hand, proposition 4.1 indicates that the conditioning of computing $(P - T_p)$ is $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$. Therefore, (42) yields that the flux Q is calculated with a conditioning of order $\mathcal{O}\left((\frac{\kappa_2}{\kappa_1})^2\right)$.

□

5 Numerical experiments

5.1 Experiment 1: stationary problem

We consider a two-dimensional elliptic boundary value problem on the unit square Ω where the boundary conditions, the permeability distribution and the value of the source term are graphically given in Fig.8. In order to flee from any numerical

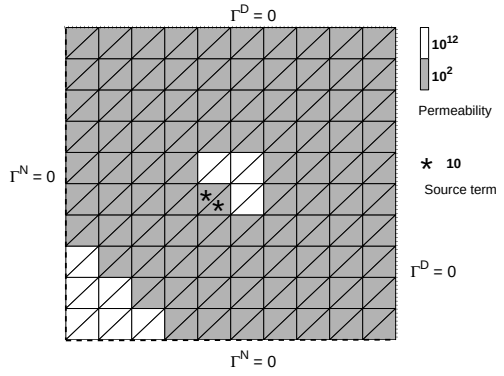

Figure 8: Triangulation of the unit square.

shortcomings that might be caused while inverting the elementary matrices, the domain is discretized into a regular grid of right angle triangular elements. In this case, the inverses of these matrices are known analytically (see [6, 14]). The solution of this problem is approximated by using both mixed methods. In the MHFE, the preconditioned conjugate gradient method is used to solve the positive definite linear system (20), whereas, in the MFE, the **Symmlq** solver (see [22, 23]) is used to solve

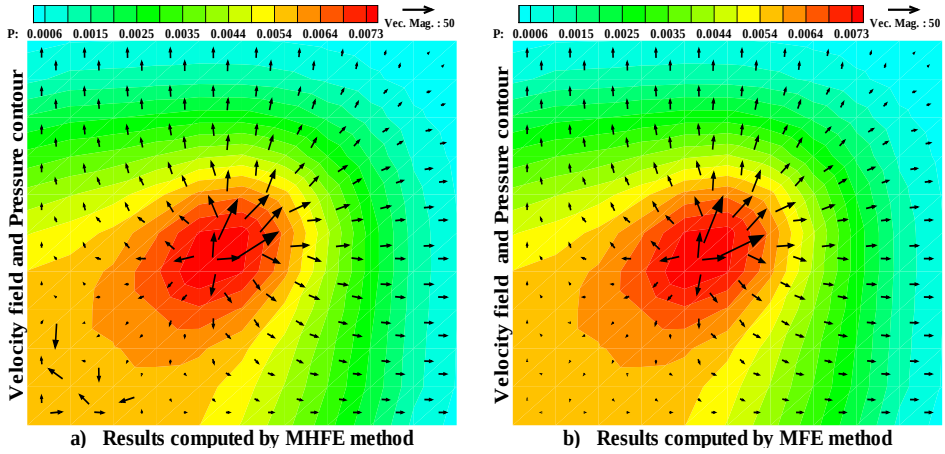

Figure 9: Numerical results computed by using MHFE and MFE codes.

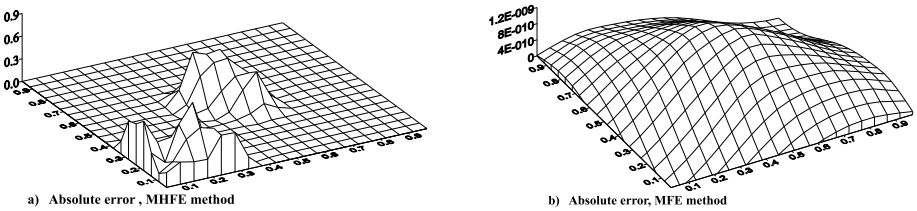

Figure 10: Absolute error obtained over the grid.

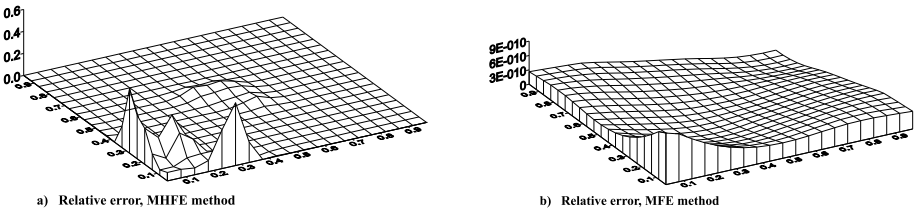

Figure 11: Relative errors obtained over the grid.

the indefinite system given in (30).
Even though the approximated pressures by both methods are nearly alike, one can

clearly notice the senseless values of the velocity field in the lower left corner of the domain (Fig.9a) which are obtained by the MHFE method. This shortcoming is due to the numerical errors indicated in (38). To evaluate the intensity of these errors, in Fig.10, we compute the absolute value of equation (40) over each element in the mesh. The relative errors depicted in Fig.11 are locally computed with respect to the sum of the absolute value of the fluxes across the element edges. In Fig.10a, we can notice that the absolute errors in the two regions of higher permeability (the two unshaded regions in Fig.8) have nearly the same order. However, the big relative errors appearing in Fig.11a, which are due to feeble values of fluxes, cause the pointless results in (Fig.9a). In contrast, the MFE is free from this difficulty.

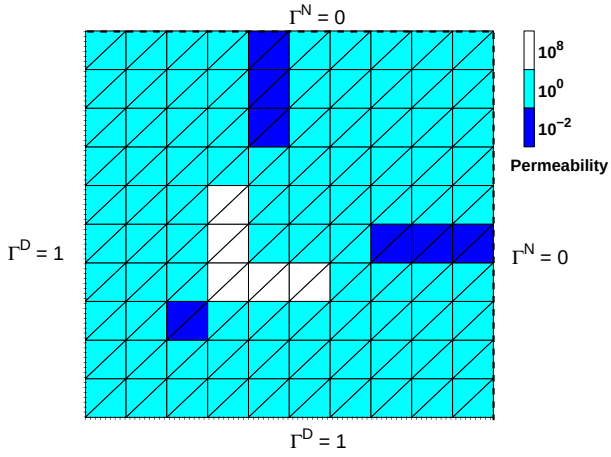

Figure 12: Permeability distribution and boundary conditions,
$T =]0, 1],\ \Delta t = \frac{1}{10},\ s = 1,\ f = 0.$

5.2 Experiment 2: parabolic problem

In Fig.13, we compare the transient solution of a parabolic problem approximated by using the two mixed methods. The simulation time interval is $]0, 1]$ with time-step $\Delta t = 1/10$, the boundary conditions and permeability values are given graphically in Fig.12. In a similar behavior as that in the elliptic case given in the former example, one can clearly notice (Fig.13a) numerical confusions in the velocity field obtained by the MHFE over the spots of high permeability. On the other hand, the MFE solution does not face this numerical difficulty (Fig.13b).

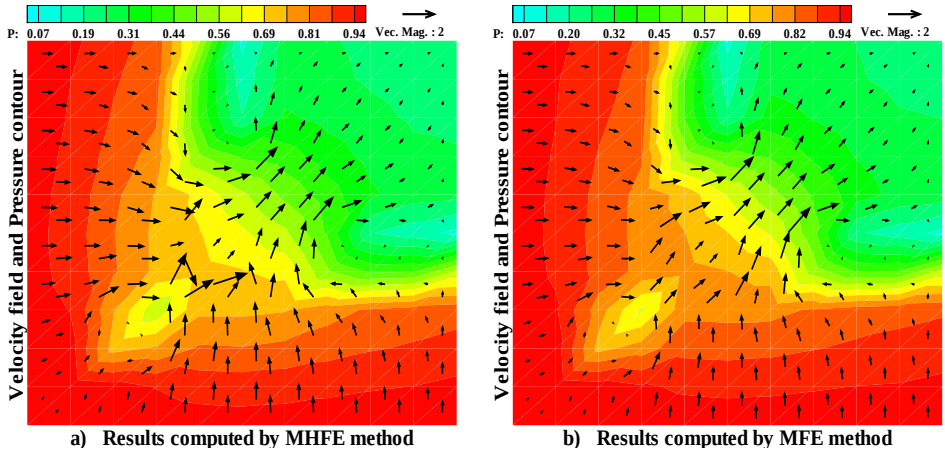

Figure 13: Numerical results computed by using MHFE and MFE codes.

5.3 Time requirements

To illustrate the theoretical results obtained in propositions 4.1, 4.2, we give some computational measurements comparing the MHFE and MFE algorithms. The domain of simulation $\Omega =]0, 20[\times]0, 20[$ is divided into two sub-domains $\Omega_2 =]5, 10[\times]5, 10[$ and $\Omega_1 = \Omega \setminus \Omega_2$. The domain is uniformly discretized into a 20×20 grid (40×40 triangular elements). Let s_i, κ_i be respectively the storage coefficient and the permeability parameter over Ω_i, $i = 1, 2$. The imposed boundary conditions are $\Gamma^D = 1$ on the left and floor sides of Ω, and $\Gamma^N = 0$ on the rest of the boundary.

The CPU running-time, the condition numbers and the number of iterations, needed to fulfill the desired termination criterion of the PCG (Preconditioned Conjugate Gradient) and **symmlq** solvers, are tabulated below. These computations were done in double precision on a Sun Ultra30 workstation. In Table 2, the domain is supposed to be homogeneous such that $s_i = \kappa_i = 1$, while we vary the time-steps. Here, the PCG algorithm is used to solve the two Schur complement systems (18), (29). In this case, the MFE algorithm is about 30% faster than the first one.
By increasing the ratio between the highest and lowest conductivities in table 3, we see that the condition numbers grow up almost linearly with the fraction $\frac{k_2}{k_1}$. Similar remarks can be noticed by using **symmlq** (see table 5) to solve the indefinite system

Table 2: $s_i = \kappa_i = 1,\ i = 1, 2,$ PCG solver.

	MHFE method			MFE method		
# Time-steps	CPU	Cond. Num.	# Iter.	CPU	Cond. Num.	# Iter
10	2.25	11.1	15–14	1.50	6.9	14–13
10^2	21.2	10.3	11–10	14.6	4.4	12–9
10^3	212.9	10.3	12–10	144.7	4.6	13–9

(28) induced by the MFE formulation.

Table 3: $\Delta t = 1/10,\ s_i = 1\ i = 1, 2,$ PCG solver.

	MHFE method			MFE method		
κ_2/κ_1	CPU	Cond. Num.	# Iter.	CPU	Cond. Num.	# Iter
10^2	2.2	10.2×10^2	22–21	1.8	4.3×10^2	33–31
10^4	2.3	10.9×10^4	24–23	2.1	4.2×10^4	42–40
10^6	2.4	10.7×10^6	26–25	2.4	4.5×10^6	55–54

Finally, as appears in table 4, the conditioning of the MFE linear system (29) grows up linearly with the quantity $\Delta t \, \|S^{-1}\|$. On the other hand, the conditionings of the MHFE system (18) (table 4) and the MFE system (28) (table 5) stay invariant as the storage coefficient tends to zero.

It should be noted that **symmlq** is used without preconditioner. The challenging point is therefore the construction of a suitable preconditioner for the indefinite system (28). This is an ongoing work.

Table 4: $\Delta t = 1/10,\ \kappa_i = 1\ i = 1, 2,$ PCG solver.

	MHFE method			MFE method		
s	CPU	Cond. Num.	# Iter.	CPU	Cond. Num.	# Iter.
10^{-2}	2.2	50.9	20–19	2.4	65.1×10^1	30–29
10^{-4}	2.2	55.1	20–19	3.2	68.3×10^3	51–50
10^{-6}	2.2	55.1	20–19	3.9	68.3×10^5	84–83

Table 5: MFE Method, **Symmlq** solver.

	$\Delta t = 1/10,\ s_i = 1\ i = 1,2.$				$\Delta t = 1/10,\ \kappa_i = 1\ i = 1,2.$		
κ_2/κ_1	CPU	Cond. Num.	# Iter.	s	CPU	Cond. Num.	# Iter.
10^2	3.9	20.2×10^2	71	10^{-2}	4.5	63.4	90
10^4	3.9	19.3×10^4	71	10^{-4}	4.7	64.8	92
10^6	4.3	19.6×10^6	75	10^{-6}	4.7	64.8	92

Conclusion

In spite of the fact that the MFE and MHFE formulations are algebraically equivalent, it is found that in many applications their numerical solutions behave differently. In this work, our ultimate intention was to check out the numerical reliability and time requirements of both algorithms under the influence of two factors: the geometry of the mesh and the medium heterogeneity. As a result, the following topmost points can be drawn.

The MHFE formulation necessitates inverting the elementary matrices. In view of the fact that flat elements could blow up the conditioning of their corresponding elementary matrices, one should be careful in choosing a stable matrix-inversion solver. While in contrast, the MFE formulation is free from this numerical difficulty since the inversion of these matrices is needless.

In heterogeneous media, it is found that the conditioning of the pressure head obtained by the MHFE method is proportional to the ratio between the highest and the lowest values of permeability parameters between adjacent subdomains, i.e. $\mathcal{O}(\frac{\kappa_2}{\kappa_1})$. Beyond, the conditioning of the computed fluxes is $\mathcal{O}((\frac{\kappa_2}{\kappa_2})^2)$. On the other hand, the MFE formulation leads to two approaches by which the unknown variables can be approximated. The first one, which is possible in the case of pure parabolic problem, is to solve the Schur complement system (18). This system is positive definite but its condition number depends on $\|S^{-1}\|$. Thus, small values of the storage coefficients have a critical effect on its resolution, commonly this is the case. The second approach leads to solve system (28) whose coefficient matrix is indefinite besides its large size compared to the first one.

Practically, the accuracy of the velocity field of a variety of groundwater flow prob-

lems, such as transport problems, is crucial. As we have seen, the MHFE formulation leads sometimes to senseless values in the velocity field, especially when large jumps in the hydraulic conductivity take place or with the presence of flat discretized elements. While on the contrary, by using the MFE method, the pressure head and the velocity field are computed with the same order of convergence. Throughout all the tested numerical experiments, no lapses in the approximated MFE solutions have been reported. As a conclusion the choice of a method is one of the cases tabulated in table 6.

Table 6: Choice between the MHFE and MFE formulations.

	$(\frac{\kappa_2}{\kappa_1})^2$ small	$\|S^{-1}\|$ small	$\|S\|$ small
MHFE	×		
MFE$^{(1)}$		×	
MFE$^{(2)}$			×

MFE$^{(1)}$: is the first approach that leads to solve the Schur complement system (29).
MFE$^{(2)}$: is the second approach which leads to solve the indefinite system (28).

The time-consuming of the MFE and MHFE algorithms strongly depends on the geometrical and physical parameters of each problem and on the linear systems to solve, so no preferences can be given for one over the other.

References

[1] G. Yeh, *Computational Subsurface Hydrology: Fluid Flows*, Kluwer Academic Publishers, The Pennsylvania State University, 1999.

[2] J. Thomas, Sur l'Analyse Numérique des Méthodes d'Elément Finis Hybrides et Mixtes *Thèse de Doctorat d'Etat*, Univ. de Pierre et Marie Curie, 1977.

[3] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation*, Elsevier Science Publishers B.V, Netherlands, 1986.

[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Method*, Springer-Verlag, New York, 1991.

[5] I. Yotov, Mixed Finite Element Methods for Flow in Porous Media, *Ph.D thesis*, Univ. of Texas, 1996.

[6] G. Chavent and J-E. Roberts, A unified physical presentation of mixed, mixed-hybrid finite element method and standard finite difference approximations for the determination of velocities in water flow problems, *Adv. Water Resour.*, 14(6) (1991) 329–348.

[7] G. Golub and C. Van Loan, *Matrix Computations*, The Johons Hopkins University Press, 1989.

[8] N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[9] J-P. Hennart, Nodal Schemes, Mixed-Hybrid finite elements and Block-Centered Finite Differences, *Rapport de Recherche*, N^o 386, INRIA, 1985.

[10] T. Russell and M. Wheeler, Finite element and finite difference methods for continuous flows in porous media, *SIAM* , R. Ewing, The Mathematics of Reservoir Simulation, (1983) 35–106 .

[11] E. Kaasschieter and A. Huijben, Mixed–Hybrid Finite Elements and streamline Computation for the Potential Flow Problem, *TNO Institute of Applied Geoscience*, TNO–Report PN 90-02-A, 1990.

[12] P. Knabner, C. Tapp and K. Thiele, Adaptive Finite Volume Discretization of Density Driven Flows in Porous Media, *Acta Mathematica Universitatis Comenianae*, (AMUC) 67 (1998) 115–136.

[13] R. Mosé, P. Siegel, Ph. Ackerer and G. Chavent, Application of the mixed hybrid finite element approximation in a groundwater flow model: Luxury or necessity?, *Water Resourc.*, 30 (1994) 3001–3012.

[14] H. Hoteit, R. Mosé, B. Philippe, Ph. Ackerer and J. Erhel, About the Maximum Principle Violations of the Mixed–Hybrid Finite Element Method applied to Diffusion Equations, (in preparation).

[15] A. Younès, R. Mose, Ph. Ackerer and G. Chavent, A New Formulation of the Mixed Finite Element Method for Solving Elliptic and Parabolic PDE with Triangular Elements, *Journal of Comp. Phys.*, 149 (1999) 148–167.

[16] Cordes and Kinzelbach, Comment on "Application of the mixed-hybrid finite approximation in a groundwater flow model: luxury or necessity, *Water Resourc.*, 32(6) (1996) 1905–1909.

[17] E. Brière and P. George, Optimization of Tetrahedral Meshes, *Springer–Verlag*, Modeling, Mesh Generation, and Adaptive Numerical Methods for PDE, 75 (1995) 97–127.

[18] J. Erhel, Experiments with Data Perturbations to Study Condition Numbers and Numerical Stability, *Computing*, 51(1) (1993) 29–44.

[19] O. Beaumont, J. Erhel, and B. Philippe, *Problem-solving environments for computational science*, to appear in IEEE Press.

[20] T. Rowan, Functional Stability Analysis of Numerical Algorithms, *Ph.D thesis*, Univ. of Texas (Austin), 1990.

[21] G. Chavent, A. Younès, R. Mosé and Ph. Ackerer, On the Finite Volume Reformation of the Mixed Finite Elements Method on Triangles, *Rapport de Recherche*, N^o 3769, INRIA, 1999.

[22] R. Barett et al., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1994.

[23] Paige and M.A. Saunders, Solution of Sparse Indefinite Systems of Linear Equations, *SIAM J. Numer. Anal.*, 12(4) (1975) 617–629.

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399